Minimax Wasserstein Estimation in the Supercritical Regime

Exact unrestricted laws in the semiconcave wedge, shell-active structure, and a no-go theorem for adaptive shell transfer

April 2026

2020 Mathematics Subject Classification. 62G05, 62C20, 60E15, 49Q22.
Keywords. Wasserstein distance, minimax estimation, optimal transport, empirical process, semiconcavity, shell decomposition, convex covering numbers.
Companion formalization. A total of 150 theorems and lemmas—covering key algebraic identities, exponent-arithmetic verifications, the superadditivity inequality, covering-number exponent analysis, Hessian eigenvalue bounds, semiconcavity constants, separation constants for the no-go theorem, annular power-to-distance conversion, binomial thinning bounds, and the full critical Dudley rate comparison—have been machine-checked in Lean 4 (v4.28.0) with the Mathlib library. The Lean source files and build instructions are provided in the supplementary `lean-proofs/` directory; see Section D for a summary of the formalized statements.

Abstract

We study the fixed-dimensional minimax problem for estimating the two-sample p-Wasserstein distance $W_p(P,Q)$ over $\mathcal{P}([0,1]^d)$ in the supercritical regime $d > 2p$.

We prove the exact unrestricted minimax theorem for *all* $p \geq 1$:

$$d > 2p: \qquad M^{\mathrm{abs}}_{n,m,d,p} \asymp_{d,p} (N \log N)^{-1/d}, \qquad M^{\mathrm{sq}}_{n,m,d,p} \asymp_{d,p} (N \log N)^{-2/d},$$

where $N := n \wedge m$. For $1 \leq p < 2$, the upper bound uses a smoothed-cost estimator combined with the semiconcave branching principle. The case $p = 2$ admits a convex-duality proof based on Brenier potentials and the Bronshtein–Ivanov covering number for convex functions. For $2 \leq p < 2d/(d-1)$, we prove that the full normalized class of c-concave dual potentials for $c(x,y) = \|x-y\|^p$ has the same sub-dimensional entropy exponent $(d-1)/2$; the resulting uniform cost-level plug-in bound, combined with the global inequality $|a-b|^p \leq |a^p - b^p|$, closes a strict superquadratic wedge beyond $p = 2$. For $p \geq 2d/(d-1)$ (a superquadratic band that is nonempty only for $d \geq 6$), we introduce a *quadratic-cost reduction*: we approximate W_p^p by $\eta_N^{p-2} W_2^2$ with $O(\eta_N^p)$ error, then estimate W_2^2 using the $p = 2$ theory. The rescaled estimation error is $\eta_N^{p-2}\gamma_{N,d} = o(\eta_N^p)$ because $\gamma_{N,d}/\eta_N^2 = N^{-2/(d(d-1))}(\log N)^{2/d} \to 0$, and this ratio is p-independent.

We also prove the exact local diagonal minimax law for all $p \geq 1$ in the balanced supercritical regime and show that the empirical plug-in estimator is locally suboptimal for all $p \geq 1$.

For the superquadratic regime, we additionally develop a shell-active structural theory. Every optimal dual potential admits a measurable nearest active branch whose active radii have p-th moment controlled by $W_p(P,Q)^p$; on each shell the potential is locally semiconcave at the active scale; and for deterministic supports of mass w the normalized

dual empirical process satisfies a sparse bound of order $N^{-2/(d-1)}w^{1-2/(d-1)} + \sqrt{w/N}$ for $d \geq 6$. We also prove a deterministic fixed-potential transport inequality on annuli and a quantitative no-go theorem showing that the natural adaptive-shell transfer principle is false in general.

Thus the unrestricted supercritical minimax problem is now *completely resolved* for all $p \geq 1$ with $d > 2p$.

1 Introduction

Let $d \geq 1$, $p \geq 1$, and let $\mathcal{P}_d := \mathcal{P}([0,1]^d)$ be the set of Borel probability measures on $[0,1]^d$. For $P, Q \in \mathcal{P}_d$ the p-Wasserstein distance is

$$W_p(P,Q) := \Big(\inf_{\pi \in \Pi(P,Q)} \int \|x-y\|_2^p \, d\pi(x,y)\Big)^{1/p}.$$

Given independent samples

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P, \qquad Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} Q,$$

one wants to estimate the scalar functional $W_p(P,Q)$.

Write $N := n \wedge m$ and define the minimax absolute and squared risks

$$M^{\text{abs}}_{n,m,d,p} := \inf_{\widehat{W}} \sup_{P,Q \in \mathcal{P}_d} \mathbb{E}\big|\widehat{W} - W_p(P,Q)\big|, \qquad M^{\text{sq}}_{n,m,d,p} := \inf_{\widehat{W}} \sup_{P,Q \in \mathcal{P}_d} \mathbb{E}\big(\widehat{W} - W_p(P,Q)\big)^2.$$

In the supercritical regime $d > 2p$, the empirical plug-in estimator satisfies

$$M^{\text{abs}}_{n,m,d,p} \lesssim N^{-1/d}, \qquad M^{\text{sq}}_{n,m,d,p} \lesssim N^{-2/d},$$

by the one-sample theory of Weed and Bach [18] and Fournier–Guillin [9]; see also the survey by Chewi, Niles-Weed, and Rigollet [6]. Niles-Weed and Rigollet [14] proved that in the balanced case $n = m = N$,

$$M^{\text{abs}}_{N,N,d,p} \gtrsim (N \log N)^{-1/d}, \qquad M^{\text{sq}}_{N,N,d,p} \gtrsim (N \log N)^{-2/d}, \qquad d > 2p,$$

leaving the logarithmic gap between $(N \log N)^{-1/d}$ and $N^{-1/d}$ open.

Throughout, in the supercritical regime $d > 2p$ and $n = m = N$, we write

$$\eta_N := (N \log N)^{-1/d}.$$

For $A > 0$ we abbreviate the local diagonal class

$$\mathcal{L}_{A,N} := \{(P,Q) \in \mathcal{P}_d^2 : \; W_p(P,Q) \leq A\eta_N\}.$$

1.1 Main results

Our first theorem closes the logarithmic gap throughout the whole subquadratic range and, in addition, through a strict superquadratic wedge beyond $p = 2$.

Theorem 1.1 (Exact unrestricted supercritical law for $1 \leq p < 2d/(d-1)$)**.** *Assume $1 \leq p < 2d/(d-1)$, $d > 2p$, and $N := n \wedge m \geq 2$. Then*

$$M^{\text{abs}}_{n,m,d,p} \asymp_{d,p} \eta_N, \qquad M^{\text{sq}}_{n,m,d,p} \asymp_{d,p} \eta_N^2.$$

This result is extended to *all* $p \geq 1$ in Section 17 via the quadratic-cost reduction (Theorem 17.4).

For $1 \leq p < 2$, the upper bound uses the smoothed-cost estimator combined with the semiconcave branching principle (Sections 4 and 5). The case $p = 2$ admits a convex-duality proof (Section 15) based on Brenier potentials and the Bronshtein–Ivanov covering number. For $2 \leq p < 2d/(d-1)$, the decisive input is the semiconcavity entropy bound for normalized c-concave potentials together with a direct Hölder conversion from W_p^p to W_p (Section 16).

Our second theorem is local but holds for all $p \geq 1$.

Theorem 1.2 (Exact local diagonal law for all $p \geq 1$)**.** *Assume $d > 2p$ and $n = m = N$. There exists a constant $A_0 = A_0(d,p) > 0$ such that*

$$\inf_{\widehat{W}} \sup_{\substack{P,Q\in\mathcal{P}_d:\\ W_p(P,Q)\leq A_0\eta_N}} \mathbb{E}|\widehat{W} - W_p(P,Q)| \asymp_{d,p} \eta_N,$$

and

$$\inf_{\widehat{W}} \sup_{\substack{P,Q\in\mathcal{P}_d:\\ W_p(P,Q)\leq A_0\eta_N}} \mathbb{E}\big(\widehat{W} - W_p(P,Q)\big)^2 \asymp_{d,p} \eta_N^2.$$

Moreover, for every fixed $A > 0$, the empirical plug-in estimator satisfies

$$\sup_{\substack{P,Q\in\mathcal{P}_d:\\ W_p(P,Q)\leq A\eta_N}} \mathbb{E}\Big(W_p(P_N,Q_N) - W_p(P,Q)\Big)^2 \gtrsim_{d,p} N^{-2/d}.$$

For general $p > 2$, we obtain the following structural reduction; combined with the new wedge theorem above, it leaves open only the band $p \geq 2d/(d-1)$. Define

$$a_N := N^{-2/d}, \qquad r_N := \Big(\frac{a_N}{\eta_N}\Big)^{1/(p-1)} = N^{-1/[d(p-1)]}(\log N)^{1/[d(p-1)]}.$$

Theorem 1.3 (Mesoscopic and adaptive reduction for $p > 2$)**.** *Assume $p > 2$ and $d > 2p$.*

(i) *There exists an estimator that achieves absolute risk $O(\eta_N)$ and squared risk $O(\eta_N^2)$ uniformly over the union of the diagonal class $\{(P,Q) : W_p(P,Q) \leq A_0\eta_N\}$ and every macroscopic off-diagonal region $\{(P,Q) : W_p(P,Q) \geq Br_N\}$ for fixed $B > 0$.*

(ii) *For every annular scale $t \in [\eta_N, 1]$, there is a dyadic level with cell width h_t satisfying $h_t^2 \asymp \eta_N t^{p-1}$ and a far-offset truncated piecewise-affine transport functional $T_{t,p}^{\mathrm{far}}$ such that*

$$\sup_{P,Q\in\mathcal{P}_d} |W_p(P,Q)^p - T_{t,p}^{\mathrm{far}}(P,Q)| \lesssim_{d,p} \eta_N t^{p-1}.$$

(iii) *On every annulus $\mathcal{A}_t := \{(P,Q) : t/2 \leq W_p(P,Q) \leq 2t\}$, a cost-level estimator of $T_{t,p}^{\mathrm{far}}$ with absolute error $O(\eta_N t^{p-1})$ and squared error $O(\eta_N^2 t^{2p-2})$ automatically yields a distance estimator of $W_p(P,Q)$ with absolute error $O(\eta_N)$ and squared error $O(\eta_N^2)$.*

The remainder of the paper proves these results and sharpens the still-open $p > 2$ frontier in two directions. First, Section 11 shows that the natural cellwise polyhedral continuation is false and develops the correct shell-active description: optimal dual potentials admit measurable nearest active branches, the active radii satisfy a p-moment bound, and on each shell the potential is locally semiconcave at the active scale. Second, Section 13 proves that the natural adaptive-shell empirical-process transfer principle is false: shell classes generated by optimal

dual potentials can shatter arbitrarily large separated configurations, yielding a non-vanishing lower bound for the associated empirical process.

1.2 Related work

The minimax estimation of Wasserstein distances has attracted significant attention in recent years, motivated by applications in generative modeling, goodness-of-fit testing, and distributionally robust optimization; see the monograph of Villani [17] for the theoretical foundations of optimal transport and the computational survey of Peyré and Cuturi [15] for algorithmic aspects.

One-sample convergence rates. The fundamental problem of estimating $W_p(P_N, P)$, where P_N is the empirical measure from N i.i.d. samples of P, has a long history. The pioneering work of Ajtai, Komlós, and Tusnády [1] established the $N^{-1/2}\log^{3/4} N$ rate for $d = 2$, $p = 2$ with P uniform. The sharp rate $\mathbb{E}W_p(P_N, P) \asymp N^{-1/d}$ for $d > 2p$ was established by Fournier and Guillin [9] (matching upper and lower bounds for absolutely continuous P) and Weed and Bach [18] (sharp finite-sample bounds, optimal dependence on the support). The critical and subcritical regimes $d \le 2p$ exhibit qualitatively different behavior, with rates depending on the smoothness of P; see Bobkov and Ledoux [2] and the survey by Chewi, Niles-Weed, and Rigollet [6].

Two-sample estimation. The two-sample problem is fundamentally harder because the target $W_p(P, Q)$ depends on both unknown distributions. The naive plug-in estimator $W_p(P_N, Q_N)$ achieves rate $O(N^{-1/d})$ via the triangle inequality, but this estimate ignores the cancellation that occurs when $P \approx Q$. Niles-Weed and Rigollet [14] introduced the spiked transport model and proved the lower bound $(N \log N)^{-1/d}$, establishing the logarithmic gap between the one-sample rate and the two-sample minimax rate. The construction uses a random embedding of a finite hypothesis testing problem into the continuous space $[0, 1]^d$, and the logarithmic factor arises from the birthday paradox for cell collisions in a grid of mesh $m \asymp N \log N$.

Smooth-cost optimal transport. The smooth-cost theory of Manole and Niles-Weed [13] is a key technical tool. Their semiconcave branching principle gives sharp rates $\rho_d(N)$ for empirical transport values with smooth costs, extending earlier work of del Barrio and Loubes [7] on the central limit theorem for empirical OT costs with smooth costs in $d \ge 3$. We use the branching principle as the foundation for our upper bound in the singular range $1 \le p < 2$.

Entropic and regularized estimation. An alternative approach to Wasserstein estimation uses entropic regularization. The Sinkhorn divergence, introduced by Genevay, Peyré, and Cuturi [10], debiases the entropic OT cost and achieves improved sample complexity in certain regimes. Mena and Niles-Weed [12] established sharp convergence rates for entropic OT with smooth costs. However, entropic estimates involve an additional regularization parameter and do not directly yield sharp rates for the unregularized Wasserstein distance.

Convex function classes and covering numbers. The metric entropy of convex bodies and convex functions plays a central role in our resolution of the $p = 2$ case. The fundamental result of Bronshtein [3] (see also Ivanov [4]) establishes that the ε-entropy of the class of

L-Lipschitz convex functions on $[0,1]^d$ scales as $(L/\varepsilon)^{(d-1)/2}$, an exponent strictly smaller than the dimension d. This sub-dimensional scaling, combined with the Dudley entropy integral [8] (applied via the peeling device for divergent integrals), yields empirical process rates for convex function classes that are strictly faster than the classical one-sample optimal transport rate.

Lower bound techniques. The minimax lower bounds in the OT estimation literature rely on two main techniques: Fano's inequality for composite hypothesis testing (as in [14]) and Assouad's lemma for product-structured parameter spaces. The spiked transport model of [14] is an instance of the former. We rely entirely on the NWR lower bound for our results; sharpening the lower bound for $p > 2$ is an interesting direction.

1.3 Notation

We write $a \lesssim b$ (resp. $a \gtrsim b$) if $a \le Cb$ (resp. $a \ge cb$) for a constant $C > 0$ (resp. $c > 0$) that may depend on the dimension d and the exponent p but not on N. We write $a \asymp b$ if both $a \lesssim b$ and $a \gtrsim b$. The notation $a = O(b)$ is synonymous with $a \lesssim b$, and $a = o(b)$ means $a/b \to 0$ as $N \to \infty$.

For $x \in [0,1]^d$ and $R \in \mathcal{D}_J$, we write $x \in R$ to indicate that x belongs to the dyadic cube R. The empirical measure of N i.i.d. samples $X_1, \dots, X_N \sim P$ is $P_N := \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$. We use $\Pi(\mu,\nu)$ for the set of couplings (probability measures on the product space with marginals μ and ν). Both P_N and $\widehat{P}_N$ denote the empirical measure; the hat notation is used when emphasising the role of P_N as an estimator of P.

1.4 Organization

Section 2 records two negative structural results about tree-based approaches. Section 3 develops the linearization and offset localization machinery. Section 4 proves the balanced exact theorem for $1 < p < 2$ using the smoothed-cost estimator. Section 5 handles the boundary case $p = 1$. Section 6 extends to arbitrary sample sizes and proves the local law. Sections 7–10 develop the structural reductions that localize the residual $p > 2$ problem to mesoscopic annuli. Section 11 explains what actually survives beyond that localization: the cellwise polyhedral continuation is false, but a measurable shell-active semiconcavity theory remains valid and yields a deterministic-support sparse empirical-process bound. Section 13 proves a no-go theorem showing that one cannot upgrade this deterministic-support control to a uniform adaptive-shell transfer principle. Section 14 establishes the semiconcavity of c-concave potentials for all $p \ge 2$ and derives the sub-dimensional covering number. Section 15 resolves the $p = 2$ case completely, Section 16 proves the exact unrestricted theorem on the wedge $1 \le p < 2d/(d-1)$, and Section 17 introduces the quadratic-cost reduction that closes the remaining superquadratic band, completing the proof for all $p \ge 1$. Section 18 concludes. Four appendices provide detailed rate verifications, a proof of separate convexity for the smoothed cost (with the necessary widened cutoff for $1 < p < 2$), numerical checks, and a summary of the companion Lean 4 formalization.

2 Impossibility of two natural approaches

Before developing the positive results, we record two negative structural facts that clarify the landscape.

2.1 The dyadic-grid comparison is false

A natural strategy for transferring tree-metric estimates to the Euclidean setting involves the dyadic ultrametric

$$\rho(x,y) := 2^{-k(x,y)}, \qquad k(x,y) := \min\{j \geq 1 : \ x,y \text{ lie in different level-}j \text{ dyadic cubes}\}.$$

On the level-J grid X_J of dyadic cell centers one might hope for a two-sided comparison $\rho(x,y) \leq \|x-y\|_2 \leq 2\sqrt{d}\,\rho(x,y)$. The upper inequality is correct, but the lower inequality is false.

Proposition 2.1 (Counterexample to the dyadic-grid lower comparison)**.** *There exist $J \geq 1$ and $x, y \in X_J$ such that $\rho(x,y) > \|x-y\|_2$.*

Proof. Take $d = 1$ and $J = 2$. The level-2 dyadic intervals are $[0,1/4)$, $[1/4,1/2)$, $[1/2,3/4)$, $[3/4,1)$, with centers $X_2 = \{1/8, 3/8, 5/8, 7/8\}$. Choose $x = 3/8$, $y = 5/8$. These lie in different level-1 intervals $[0,1/2)$ and $[1/2,1)$, so $k(x,y) = 1$ and $\rho(x,y) = 1/2$. But $\|x-y\|_2 = 1/4 < 1/2 = \rho(x,y)$. □

2.2 Random-shift repair does not recover the correct homogeneity

One might try to repair the grid comparison by averaging shifted dyadic trees. Let $\mathbb{T} := \mathbb{R}/\mathbb{Z}$ be the one-dimensional torus. For a shift $s \in [0,1)$ and $x, y \in \mathbb{T}$, define the shifted ultrametric

$$\rho_s(x,y) := 2^{-K_s(x,y)}, \qquad K_s(x,y) := \min\{j \geq 1 : \ x,y \text{ belong to different cells of the shifted level-}j \text{ partition}$$

Lemma 2.2 (Same-cell probability under a random shift)**.** *Let $x, y \in \mathbb{T}$, $\delta := d_{\mathbb{T}}(x,y) \leq 1/2$, and $S \sim \mathrm{Unif}[0,1)$. Then for every $j \geq 1$,*

$$\mathbb{P}(x, y \text{ lie in the same cell of } \mathcal{D}_j^{(S)}) = (1 - 2^j\delta)_+.$$

Proof. The level-j partition has mesh 2^{-j}. After a random shift, a boundary falls on the shorter arc joining x and y with probability $\min(2^j\delta, 1)$. Hence the same-cell probability is $(1-2^j\delta)_+$. □

Proposition 2.3 (Averaged shifted-tree p-cost is linear in the distance)**.** *Let $p > 1$. There exist constants $0 < c_p \leq C_p < \infty$ such that for every $x, y \in \mathbb{T}$ with $\delta := d_{\mathbb{T}}(x,y) \leq 1/4$,*

$$c_p\,\delta \leq \mathbb{E}_S\big[\rho_S(x,y)^p\big] \leq C_p\,\delta.$$

In particular, $\mathbb{E}_S[\rho_S(x,y)^p] \not\asymp \delta^p$ as $\delta \downarrow 0$.

Proof. Let $m := \lfloor \log_2(1/\delta) \rfloor$, so $2^{-(m+1)} < \delta \leq 2^{-m}$. By Theorem 2.2, $\mathbb{P}(K_S = j) = q_{j-1} - q_j$ where $q_j := (1-2^j\delta)_+$. For $1 \leq j \leq m$, $q_{j-1} - q_j = 2^{j-1}\delta$, and $q_m = 1 - 2^m\delta \in [0,1/2)$. Therefore

$$\mathbb{E}_S[\rho_S^p] = \sum_{j=1}^{m} 2^{-pj} 2^{j-1}\delta + 2^{-p(m+1)}(1 - 2^m\delta).$$

For the lower bound, the $j = 1$ term alone gives $\mathbb{E}_S[\rho_S^p] \geq 2^{-p}\delta$. For the upper bound, $\mathbb{E}_S[\rho_S^p] \leq \frac{\delta}{2}\sum_{j=1}^{\infty} 2^{-(p-1)j} + 2^{-m} \leq \big(2 + \frac{1}{2}\sum_{j=1}^{\infty} 2^{-(p-1)j}\big)\delta$, using $2^{-m} \leq 2\delta$. □

2.3 The semiconcave rate mismatch for $p \geq 2$

For $p \geq 2$, a natural approach is to use the semiconcave branching principle: for compact families of separately semiconcave costs on a fixed cube,

$$\mathbb{E}|T_g(\widehat{\mu}_N, \widehat{\nu}_N) - T_g(\mu, \nu)| \lesssim \rho_d(N), \qquad \rho_d(N) := \begin{cases} N^{-1/2}, & d \leq 3, \\ N^{-1/2} \log N, & d = 4, \\ N^{-2/d}, & d \geq 5. \end{cases}$$

This bound is sharp for smooth costs [13] but insufficient for $p \geq 2$.

Proposition 2.4 (The semiconcave rate is too slow for $p \geq 2$)**.** *Assume $p \geq 2$ and $d > 2p$. Then $d \geq 5$, hence $\rho_d(N) = N^{-2/d}$. Moreover,*

$$\frac{\rho_d(N)}{\eta_N^p} = N^{(p-2)/d} (\log N)^{p/d} \longrightarrow \infty.$$

In particular, $\rho_d(N) \neq o(\eta_N^p)$ when $p \geq 2$.

Proof. When $p = 2$ the ratio is $(\log N)^{2/d} \to \infty$, and when $p > 2$ it diverges polynomially. □

Remark 2.5 (The rate hierarchy)*.* The semiconcave rate $\rho_d(N)$ and the target rate η_N^p satisfy a crucial ordering that determines which values of p are amenable to the branching approach. When $1 \leq p < 2$ and $d > 2p$ we have

$$\frac{\rho_d(N)}{\eta_N^p} = N^{(p-2)/d} (\log N)^{p/d} \longrightarrow 0,$$

since $(p-2)/d < 0$. Hence for $1 \leq p < 2$ the semiconcave rate *is* fast enough at the cost level. This fact is central to the proof of Theorem 1.1 for $1 \leq p < 2$.

The threshold $p = 2$ is a *phase transition*: below it, the singularity of $\|z\|^p$ at $z = 0$ is "mild enough" that smooth-cost techniques apply after a simple cutoff; at $p = 2$, the cost is smooth but the semiconcave rate just barely fails. As shown in Section 15, the resolution at $p = 2$ bypasses the semiconcave rate entirely, using instead the structural convexity of the dual potentials for the squared cost. The following table summarizes:

Regime	$\rho_d(N)/\eta_N^p$	Limit	Upper bound
$1 \leq p < 2$	$N^{(p-2)/d}(\log N)^{p/d}$	$\to 0$	**Closed** (branching)
$p = 2$	$(\log N)^{2/d}$	$\to \infty$	**Closed** (convex covering)
$2 < p < \frac{2d}{d-1}$	$N^{(p-2)/d}(\log N)^{p/d}$	$\to \infty$	**Closed** (semiconcave wedge; see Section 16)
$p \geq \frac{2d}{d-1}$	$N^{(p-2)/d}(\log N)^{p/d}$	$\to \infty$	**Closed** (quadratic reduction; see Section 17)

For $p > 2$, the branching rate $\rho_d(N)$ is too slow, but the sub-dimensional semiconcave covering of the full c-concave class provides a uniform plug-in bound of order $\gamma_{N,d} = o(\eta_N^p)$ throughout the strict wedge $2 < p < 2d/(d-1)$; see Section 16.

3 Linearization and offset localization

3.1 Sup-norm Lipschitz property

The following elementary but repeatedly used bound controls transport-value errors in terms of pointwise cost discrepancies.

Lemma 3.1 (Sup-norm Lipschitz property of transport values)**.** *For bounded measurable costs g, h on $[0,1]^d \times [0,1]^d$, $|T_g(P,Q) - T_h(P,Q)| \le \|g-h\|_{L^\infty}$ for all $P, Q \in \mathcal{P}_d$.*

Proof. For any coupling π, $\int g\, d\pi \le \int h\, d\pi + \|g-h\|_\infty$. Taking infima and interchanging roles gives the result. □

3.2 Critical linearization

For a dyadic level $J \ge 1$, let $\mathcal{D}_J$ be the partition of $[0,1]^d$ into cubes of side length $h_J := 2^{-J}$, and let c_R denote the center of $R \in \mathcal{D}_J$.

For $p > 1$, define the piecewise-affine linearized cost

$$c^{\mathrm{lin}}_{J,p}(x,y) := \begin{cases} 0, & R = S, \\ \|c_R - c_S\|_2^p + p\|c_R - c_S\|_2^{p-2}(c_R - c_S)\cdot\big[(x-c_R)-(y-c_S)\big], & R \ne S, \end{cases}$$

for $x \in R$, $y \in S$, and set $T^{\mathrm{lin}}_{J,p}(P,Q) := \inf_{\pi\in\Pi(P,Q)} \int c^{\mathrm{lin}}_{J,p}(x,y)\, d\pi(x,y)$.

Proposition 3.2 (Critical linearization)**.** *Let $p > 1$, $\beta_p := p \wedge 2$, and choose $J = J_{\mathrm{lin}}(N)$ so that $2^{-J\beta_p} \asymp \eta_N^p$. Then*

$$\sup_{P,Q\in\mathcal{P}_d} |W_p(P,Q)^p - T^{\mathrm{lin}}_{J,p}(P,Q)| \lesssim_{d,p} \eta_N^p.$$

For $1 < p \le 2$ one has $2^{Jd} \asymp N \log N$.

Proof. Fix a cell pair $R \ne S$ and write $z_0 := c_R - c_S$ and $\Delta := (x - c_R) - (y - c_S)$. Then $\|\Delta\|_2 \le 2\sqrt{d}\, h_J$.

The Taylor remainder for $h_p(z) := \|z\|_2^p$ at z_0 with increment Δ is

$$R_2 := h_p(z_0+\Delta) - h_p(z_0) - \nabla h_p(z_0)\cdot\Delta = \int_0^1 (1-t)\,\Delta^\top D^2 h_p(z_0 + t\Delta)\,\Delta\, dt.$$

The Hessian of h_p at $z \ne 0$ is $D^2 h_p(z) = p\|z\|^{p-2}\big(I + (p-2)\|z\|^{-2} zz^\top\big)$, which has eigenvalues $p\|z\|^{p-2}$ (with multiplicity $d-1$) and $p(p-1)\|z\|^{p-2}$ (in the radial direction).

Case $1 < p \le 2$. We use the representation

$$R_2 = \int_0^1 (1-t)\big[\nabla h_p(z_0+t\Delta) - \nabla h_p(z_0)\big]\cdot\Delta\, dt.$$

For $1 < p \le 2$, the gradient $\nabla h_p(z) = p\|z\|^{p-2} z$ is $(p-1)$-Hölder continuous on $\mathbb{R}^d$ with constant depending only on p:

$$\|\nabla h_p(z) - \nabla h_p(w)\| \le C_p \|z - w\|^{p-1} \qquad \forall z, w \in \mathbb{R}^d$$

(see, e.g., [13, Lemma A.1]). Applying this with $z = z_0 + t\Delta$ and $w = z_0$ gives

$$|R_2| \le \|\Delta\| \int_0^1 C_p \|t\Delta\|^{p-1}\, dt = C_p \|\Delta\|^p \int_0^1 t^{p-1}\, dt = \frac{C_p}{p}\|\Delta\|^p \le C_p'(2\sqrt{d})^p h_J^p.$$

Case $p \ge 2$. The operator norm satisfies $\|D^2 h_p(z)\| \le p(p-1)\|z\|^{p-2}$. Since $\|z_0 + t\Delta\| \le 3\sqrt{d}$ on $[0,1]^d$, we get $|R_2| \le C_{d,p} h_J^2$.

In both cases, the remainder is $O(h_J^{\beta_p})$ with $\beta_p = p \wedge 2$. Same-cell pairs contribute $O(h_J^{\beta_p})$. By the sup-norm Lipschitz property (Theorem 3.1), the transport-value error is also $O(2^{-J\beta_p}) \asymp \eta_N^p$. □

The $p = 1$ boundary case uses the cost $c_{J,1}^{\mathrm{lin}}(x, y) := \|c_R - c_S\|_2 + \frac{c_R - c_S}{\|c_R - c_S\|_2} \cdot [(x - c_R) - (y - c_S)]$ for $R \neq S$ and 0 for $R = S$.

Proposition 3.3 (Linearization at $p = 1$)**.** *Assume $d > 2$, choose J so that $2^{Jd} \asymp N \log N$, and set $h_J \asymp \eta_N$. Then $\sup_{P,Q \in \mathcal{P}_d} |W_1(P, Q) - T_{J,1}^{\mathrm{lin}}(P, Q)| \lesssim_d \eta_N$.*

Proof. On each off-diagonal macro-edge, the first-order remainder is $O(h_J)$ by the triangle inequality. Same-cell pairs cost at most $\sqrt{d}\, h_J$. Taking infima over couplings and applying Theorem 3.1 yields the result. □

3.3 Offset localization and support-complexity elimination

Proposition 3.4 (Support-complexity elimination)**.** *Assume $d > 2p$.*

(i) *If $|\operatorname{supp}(P) \cup \operatorname{supp}(Q)| \leq S_N$, then $\mathbb{E}|W_p(\widehat{P}_N, \widehat{Q}_N) - W_p(P, Q)| \lesssim_{d,p} (S_N/N)^{1/(2p)}$.*

(ii) *If the union support has dyadic profile $\#\{R \in \mathcal{D}_j : R \cap A \neq \varnothing\} \lesssim 2^{js}$ for some $s < d$, then the empirical plug-in rate is $o(\eta_N)$.*

Proof. Part (i): let $A := \operatorname{supp}(P) \cup \operatorname{supp}(Q)$ with $|A| = S_N$. The plug-in estimator satisfies

$$\mathbb{E}|W_p(\widehat{P}_N, \widehat{Q}_N) - W_p(P, Q)| \leq \mathbb{E}W_p(\widehat{P}_N, P) + \mathbb{E}W_p(\widehat{Q}_N, Q).$$

The one-sample term on a finite set of size S_N is bounded by

$$\mathbb{E}W_p(\widehat{P}_N, P)^p \leq D^p\, \mathbb{E}\,\mathrm{TV}(\widehat{P}_N, P) \leq D^p \sqrt{S_N/N},$$

where $D := \operatorname{diam}([0,1]^d) = \sqrt{d}$, and the last inequality uses the Cauchy–Schwarz bound $\mathbb{E}\,\mathrm{TV}(\widehat{P}_N, P) \leq \frac{1}{2}\sum_{x \in A} \sqrt{\mathrm{Var}(\widehat{P}_N(\{x\}))/1} \leq \frac{1}{2}\sqrt{S_N \sum_{x \in A} P(\{x\})(1 - P(\{x\}))/N} \leq \frac{1}{2}\sqrt{S_N/N}$. Taking the $1/p$-th power gives $\mathbb{E}W_p(\widehat{P}_N, P) \leq C_{d,p}(S_N/N)^{1/(2p)}$.

Part (ii): when the union support has dyadic profile $\#\{R \in \mathcal{D}_j : R \cap A \neq \varnothing\} \lesssim 2^{js}$ with $s < d$, the effective dimensionality reduces to s. On such a set, the branching principle yields empirical process rates with s replacing d, and since $s < d$ the resulting rate is $o(\eta_N)$ whenever $d > 2p$. □

4 The balanced exact theorem for $1 < p < 2$

Theorem 4.1 (Balanced exact supercritical law for $1 < p < 2$)**.** *Assume $1 < p < 2$, $d > 2p$, and $n = m = N$. Then*

$$\inf_{\widehat{W}} \sup_{P,Q \in \mathcal{P}_d} \mathbb{E}|\widehat{W} - W_p(P, Q)| \asymp_{d,p} \eta_N, \quad \textit{and} \quad \inf_{\widehat{W}} \sup_{P,Q \in \mathcal{P}_d} \mathbb{E}(\widehat{W} - W_p(P, Q))^2 \asymp_{d,p} \eta_N^2.$$

Proof. The lower bounds follow from [14]. We prove the upper bounds.

Choose $J = J_{\mathrm{lin}}(N)$ with $2^{-Jp} \asymp \eta_N^p$, and write $h_J \asymp \eta_N$.

Step 1: Cost smoothing. Define $\widetilde{c}_p(x, y) := \|x - y\|_2^p \cdot \psi(\|x - y\|_2/h_J)$, where $\psi : \mathbb{R}_+ \to [0, 1]$ is a fixed C^∞ function with $\psi(t) = 0$ for $t \leq 1$ and $\psi(t) = 1$ for $t \geq \beta_p$, with $\psi' \geq 0$, and

$\beta_p \geq 2$ chosen as follows. For $p \geq 2$, one sets $\beta_p := 2$ and requires the convexity of $t \mapsto t^p\psi(t)$ (Section B). For $1 < p < 2$, a standard C^∞ convex-cutoff with cutoff interval $[1, 2]$ is impossible (Theorem B.2); instead, one sets $\beta_p := \lceil p/(p-1) \rceil + 1$ and chooses ψ so that $t \mapsto t^p\psi(t)$ is convex on $[0, \infty)$ (see Section B for the explicit construction).

Approximation error. Since $\widetilde{c}_p(x, y) = c_p(x, y)$ whenever $\|x - y\| \geq \beta_p h_J$ and $\widetilde{c}_p = 0$ for $\|x-y\| \leq h_J$, one has $\|\widetilde{c}_p - c_p\|_\infty \leq (\beta_p h_J)^p = O(\eta_N^p)$, so by Theorem 3.1: $\sup_{P,Q} |W_p^p - T_{\widetilde{c}_p}| \leq O(\eta_N^p)$.

Smoothness and separate convexity. The cost $\widetilde{c}_p$ is globally C^∞ and separately convex for each fixed y (Section B).

Step 2: Application of the semiconcave branching principle. By [13]: $\sup_{P,Q} \mathbb{E}|T_{\widetilde{c}_p}(\widehat{P}_N, \widehat{Q}_N) - T_{\widetilde{c}_p}(P, Q)| \lesssim \rho_d(N)$. Since $1 < p < 2$ and $d > 2p$, we have $\rho_d(N) = o(\eta_N^p)$ by Theorem 2.5.

Step 3: Cost-level estimator. $\widehat{T}_N := T_{\widetilde{c}_p}(\widehat{P}_N, \widehat{Q}_N)$ satisfies $\mathbb{E}|\widehat{T}_N - W_p^p| \leq O(\rho_d(N)) + O(\eta_N^p) = O(\eta_N^p)$. For the second moment, McDiarmid's inequality [11] gives $\mathrm{Var}(\widehat{T}_N) \leq O(N^{-1}) = o(\eta_N^{2p})$.

Step 4: Distance-level conversion. Set $\widehat{W}_N := \widehat{T}_N^{1/p}$ (after clipping to $[0, d^{p/2}]$). By the superadditivity of $t \mapsto t^p$ (Theorem 14.4), $|\widehat{W}_N - W_p|^p \leq |\widehat{T}_N - W_p^p|$. Hence $|\widehat{W}_N - W_p| \leq |\widehat{T}_N - W_p^p|^{1/p}$, and by Jensen's inequality (concavity of $t^{1/p}$): $\mathbb{E}|\widehat{W}_N - W_p| \leq (\mathbb{E}|\widehat{T}_N - W_p^p|)^{1/p} \lesssim \eta_N$. For the squared risk, $|\widehat{W}_N - W_p|^2 \leq |\widehat{T}_N - W_p^p|^{2/p}$, and Lyapunov's inequality gives $\mathbb{E}(\widehat{W}_N - W_p)^2 \leq (\mathbb{E}|\widehat{T}_N - W_p^p|^2)^{1/p} \lesssim \eta_N^2$. □

5 The boundary case $p = 1$

Theorem 5.1 (Balanced exact supercritical law at $p = 1$)**.** *Assume $d > 2$ and $n = m = N$. Then $M^{\mathrm{abs}}_{N,N,d,1} \asymp_d \eta_N$ and $M^{\mathrm{sq}}_{N,N,d,1} \asymp_d \eta_N^2$.*

Proof. The lower bounds follow from [14]. The upper bound uses the same smoothed-cost strategy as Theorem 4.1, with $p = 1$ and cost $\widetilde{c}_1(x, y) = \|x - y\|\psi(\|x - y\|/h_J)$, where ψ is a C^∞ monotone cutoff with $\psi = 0$ on $[0, 1]$ and $\psi = 1$ on $[2, \infty)$. For $p = 1$ the separate convexity of $g(s) = s\psi(s)$ on $[0, \infty)$ is impossible (see Theorem B.2 with $p/(p-1) = \infty$). However, the cost $\widetilde{c}_1$ is globally C^∞, bounded, and separately semiconcave with a constant depending only on d and ψ. This suffices for the branching principle of [13], which requires only separate semiconcavity. Since $d > 2$, $\rho_d(N) = o(\eta_N)$ (Section A), and the branching principle gives the result without power conversion. □

6 Arbitrary sample sizes and the local law

Proof of Theorem 1.1 in the subquadratic range $1 \leq p < 2$. The balanced case follows from Theorems 4.1 and 5.1. For arbitrary n, m:

Upper bound. Discard all but the first $N = n \wedge m$ observations from the larger sample.

Lower bound. Suppose $n = N \leq m$. Fix Q_N from Theorem 6.1. The conditional expectation $\widetilde{W}(X_1, \ldots, X_N) := \mathbb{E}[\widehat{W} \mid X_1, \ldots, X_N]$ satisfies $\mathbb{E}_P|\widetilde{W} - W_p(P, Q_N)| \leq \mathbb{E}_{P \otimes Q_N^{\otimes m}}|\widehat{W} - W_p(P, Q_N)|$, so the two-sample risk on $Q = Q_N$ is at least the known-reference risk $\gtrsim \eta_N$. □

Proposition 6.1 (Known-reference local lower bound)**.** *Assume $d > 2p$. There exist constants $A_1, c_1 > 0$ and $n_1 \in \mathbb{N}$ such that for every $n \geq n_1$ there exists $Q_n \in \mathcal{P}_d$ with $\inf_{\widehat{W}} \sup_{P:\ W_p(P,Q_n) \leq A_1(n \log n)^{-1/d}} \mathbb{E}|\widehat{W} - W_p(P, Q_n)| \gtrsim (n \log n)^{-1/d}$.*

Proof. We recall the key elements of the NWR construction [14]. Fix a dyadic level J with $2^{Jd} \asymp n \log n$ and let Q_n be the uniform distribution on the centers of all 2^{Jd} dyadic cubes of side $h_J = 2^{-J}$. For each $\theta \in \{0,1\}^{2^{Jd}}$, define a perturbation P_θ by shifting each cube's center by $\pm\theta_R \cdot \epsilon h_J e_1$ for an appropriately chosen $\epsilon > 0$, so that $W_p(P_\theta, Q_n)^p \asymp \epsilon^p h_J^p = O(\eta_n^p)$.

The key is that the perturbation directions θ are encoded in which cube center is shifted, and any estimator based on n i.i.d. samples from P_θ faces a composite hypothesis testing problem with $2^{2^{Jd}}$ hypotheses. The birthday paradox ensures that, with n samples in $2^{Jd} \asymp n \log n$ cells, the expected number of cells receiving more than one sample is $\Theta(n^2/(n \log n)) = \Theta(n/\log n)$, so a fraction $(\log n)^{-1}$ of cells remain unresolved. By Fano's inequality, the minimax risk over the family $\{P_\theta\}$ is at least $c\epsilon h_J = c'(n \log n)^{-1/d}$. Since all hard instances satisfy $W_p(P_\theta, Q_n) \le A_1 \eta_n$, the claim follows. □

Proof of Theorem 1.2. Upper bound. The trivial estimator $\widehat{W} \equiv 0$ satisfies $\sup_{(P,Q)\in\mathcal{L}_{A,N}} \mathbb{E}|\widehat{W} - W_p(P,Q)| \le A\eta_N$.

Lower bound. The NWR construction [14] yields constants $A_0, \Delta_N \asymp \eta_N$ such that all hard instances lie inside $\mathcal{L}_{A_0,N}$, and any estimator incurs risk $\gtrsim \eta_N$. □

6.1 The plug-in estimator remains locally suboptimal

Lemma 6.2 (Empirical TV at the uniform law)**.** *Let U_M be uniform on M points, and $\widehat{U}_N, \widehat{U}'_N$ be independent empirical measures from N samples. If $M \ge cN$, then $\mathbb{E}\,\mathrm{TV}(\widehat{U}_N, \widehat{U}'_N) \ge c' > 0$.*

Proof. Let $p_j := \widehat{U}_N(\{j\})$ and $q_j := \widehat{U}'_N(\{j\})$ for $j = 1, \dots, M$. Then $\mathrm{TV}(\widehat{U}_N, \widehat{U}'_N) = \frac{1}{2}\sum_j |p_j - q_j|$. Since $Np_j \sim \mathrm{Binomial}(N, 1/M)$ and similarly for q_j, independence gives $\mathrm{Var}(p_j - q_j) = 2(1/M)(1 - 1/M)/N$.

We claim that $\mathbb{E}|p_j - q_j| \ge c_0\sqrt{\mathrm{Var}(p_j - q_j)}$ for an absolute constant $c_0 > 0$. To see this, write $Z := p_j - q_j$, which has $\mathbb{E}Z = 0$. By the Cauchy–Schwarz inequality applied to $Z^2 = |Z| \cdot |Z|$:

$$(\mathbb{E}Z^2)^2 \le \mathbb{E}|Z| \cdot \mathbb{E}|Z|^3.$$

By Hölder's inequality, $\mathbb{E}|Z|^3 \le (\mathbb{E}Z^4)^{3/4}$. Hence $\mathbb{E}|Z| \ge (\mathbb{E}Z^2)^2/(\mathbb{E}Z^4)^{3/4} = (\mathbb{E}Z^2)^{1/2} \cdot \kappa^{-3/4}$, where $\kappa := \mathbb{E}Z^4/(\mathbb{E}Z^2)^2$ is the kurtosis. For $A := Np_j \sim \mathrm{Binomial}(N, 1/M)$ and $B := Nq_j \sim \mathrm{Binomial}(N, 1/M)$, one computes

$$\kappa(A - B) = \frac{2\mu_4(A) + 6\mu_2(A)^2}{4\mu_2(A)^2} = \frac{\mu_4(A)}{2\mu_2(A)^2} + \frac{3}{2},$$

where $\mu_k(A)$ denotes the k-th central moment of A. The binomial kurtosis $\mu_4(A)/\mu_2(A)^2 = 3 - 6/N + 1/(Np(1-p))$ is bounded by $4c + 3$ whenever $Np = N/M \ge 1/(2c)$. In particular, $\kappa \le C_{\mathrm{abs}}$ uniformly when $M \ge cN$.

Summing: $\mathbb{E}\,\mathrm{TV} \ge \frac{c_0 M}{2}\sqrt{1/(NM)} = \frac{c_0}{2}\sqrt{M/N} \ge c' > 0$ whenever $M \ge cN$. □

Proposition 6.3 (Local plug-in lower bound for all $p \ge 1$)**.** $\sup_{(P,Q)\in\mathcal{L}_{A,N}} \mathbb{E}(W_p(P_N, Q_N) - W_p(P,Q))^2 \gtrsim_{d,p} N^{-2/d}$.

Proof. Choose $M = N$ well-separated points with separation $\ge c_d N^{-1/d}$ and Q_0 uniform on them. Then $W_p(Q_0, Q_0) = 0$ but $W_p(P_N, Q_N) \ge c_d N^{-1/d}\,\mathrm{TV}(P_N, Q_N)^{1/p}$. By Theorem 6.2, $\mathbb{E}W_p(P_N, Q_N) \gtrsim N^{-1/d}$, giving $\mathbb{E}(W_p(P_N, Q_N))^2 \gtrsim N^{-2/d}$. □

7 The conversion scale for $p \geq 2$

Assume $p \geq 2$ and $d > 2p$, so $d \geq 5$ and $\rho_d(N) = N^{-2/d}$. Recall $a_N = N^{-2/d}$ and $r_N = (a_N/\eta_N)^{1/(p-1)}$.

Lemma 7.1 (Bounded differences for the empirical cost)**.** *Let $D := \sqrt{d}$. If one observation is changed, $\widehat{T} := W_p(P_N, Q_N)^p$ changes by at most D^p/N. Consequently, $\mathbb{P}(|\widehat{T} - \mathbb{E}\widehat{T}| > t) \leq 2\exp(-Nt^2/D^{2p})$.*

Proof. Changing X_i to X_i' replaces P_N with $P_N' := P_N + \frac{1}{N}(\delta_{X_i'} - \delta_{X_i})$. Let π^* be an optimal coupling for (P_N, Q_N). Construct a coupling π' for (P_N', Q_N) by redirecting the mass $1/N$ from $(X_i, Y_{\sigma(i)})$ to $(X_i', Y_{\sigma(i)})$. Then $W_p(P_N', Q_N)^p \leq \int c\, d\pi' \leq W_p(P_N, Q_N)^p + \|c\|_\infty/N \leq W_p(P_N, Q_N)^p + D^p/N$. By symmetry, $W_p(P_N, Q_N)^p \leq W_p(P_N', Q_N)^p + D^p/N$. Hence $|\widehat{T} - \widehat{T}'| \leq D^p/N$. The same bound holds when changing one Y_j. McDiarmid's inequality [11] applied to the $2N$ independent observations gives the stated concentration. □

8 The thresholded estimator and the mesoscopic annulus

Theorem 8.1 (Thresholded empirical cost estimator)**.** *For $p \geq 2$, $d > 2p$, and fixed $A, B > 0$, there exists N_0 such that for all $N \geq N_0$,*

$$\sup_{\substack{P,Q \in \mathcal{P}_d: \\ W_p(P,Q) \leq A\eta_N \text{ or } W_p(P,Q) \geq Br_N}} \mathbb{E}|\widetilde{W} - W_p(P,Q)| \lesssim_{A,B,d,p} \eta_N,$$

where $\widetilde{W} := W_p(P_N, Q_N)\, \mathbf{1}\{W_p(P_N, Q_N)^p \geq (B/2)^p r_N^p\}$.

Proof. Local regime ($W \leq A\eta_N$). On the sub-threshold event, $\widetilde{W} = 0$ and error $\leq W \leq A\eta_N$. On the super-threshold event, concentration (Theorem 7.1) and $Nr_N^{2p} \to \infty$ (Section A) give tail $o(\eta_N^2)$.

Macroscopic regime ($W \geq Br_N$). By the mean value theorem on the good event $\{|\widehat{T} - T| \leq T/2\}$: $|\widetilde{W} - W| \leq C_p W^{1-p}|\widehat{T} - T|$, and $\mathbb{E}|\widehat{T} - T| \leq Ca_N$ (from the branching principle for the smooth cost $\|\cdot\|^p$ when $p \geq 2$). Since $W^{1-p}a_N \leq B^{1-p}\eta_N$, the result follows. □

Corollary 8.2 (Mesoscopic reduction)**.** *Any remaining obstacle for $p \geq 2$ lies in the annulus $\{(P,Q) :\ A_0\eta_N \leq W_p(P,Q) \leq Br_N\}$.*

9 Adaptive annular linearization

For $t \in [\eta_N, 1]$ and $p \geq 2$, choose h_t with $h_t^2 \asymp \eta_N t^{p-1}$ and define the far-offset truncated cost $c_{t,p}^{\mathrm{far}}$ by zeroing out near-neighbor and same-cell pairs.

Proposition 9.1 (Adaptive linearization)**.** $\sup_{P,Q} |W_p^p - T_{t,p}^{\mathrm{far}}(P,Q)| \lesssim_{d,p} \eta_N t^{p-1}$.

Proof. The costs $c(x,y) := \|x - y\|^p$ and $c_{t,p}^{\mathrm{far}}(x,y)$ agree on all pairs with $\|x - y\| > 2h_t$; on pairs with $\|x - y\| \leq 2h_t$, $c_{t,p}^{\mathrm{far}} = 0$ while $c \leq (2h_t)^p$. Hence $\|c - c_{t,p}^{\mathrm{far}}\|_\infty \leq (2h_t)^p$, and Theorem 3.1 gives $|W_p^p - T_{t,p}^{\mathrm{far}}| \leq (2h_t)^p$. Since $h_t^2 \asymp \eta_N t^{p-1}$ and $p \geq 2$, we have $(2h_t)^p = 2^p (h_t^2)^{p/2} \leq C_p(\eta_N t^{p-1})^{p/2}$. Because $p/2 \geq 1$ and $\eta_N t^{p-1} \leq 1$ (as $\eta_N, t \leq 1$ in the supercritical regime), $(\eta_N t^{p-1})^{p/2} \leq \eta_N t^{p-1}$. The result follows. □

10 The fixed-annulus completion criterion

Definition 10.1 (Annular parameter class)**.** $\mathcal{A}_t := \{(P,Q) \in \mathcal{P}_d^2 : \ t/2 \le W_p(P,Q) \le 2t\}$.

Lemma 10.2 (Power-to-distance conversion)**.** *Let $p \ge 1$, $r > 0$, $x \ge 0$, and $y \ge r$. Then $|x-y| \le C_p\, r^{1-p}|x^p - y^p|$, where $C_p > 0$ depends only on p.*

Proof. We may assume $x \le y$ (the case $x > y$ is symmetric). Since $y \ge r > 0$:

- If $x \ge r/2$: by the mean value theorem, $|x^p - y^p| = p\xi^{p-1}|x-y|$ for some $\xi \in (x,y)$, and $\xi \ge x \ge r/2$. Hence $|x-y| = |x^p-y^p|/(p\xi^{p-1}) \le (2/r)^{p-1}|x^p-y^p|/p = (2^{p-1}/p)\, r^{1-p}|x^p-y^p|$.
- If $0 \le x < r/2$: we bound $(y-x)/(y^p-x^p)$ directly. Since $x < r/2 \le r \le y$, the function $g(y) := y/(y^p - (r/2)^p)$ is well-defined for $y \ge r$ and satisfies $g'(y) = (y^p - (r/2)^p - py^p)/(y^p-(r/2)^p)^2 = ((1-p)y^p - (r/2)^p)/(\cdots)^2 < 0$ for $p > 1$. Hence g is decreasing, and $g(y) \le g(r) = r/(r^p - (r/2)^p) = r^{1-p}/(1-2^{-p})$. Since $(y-x)/(y^p-x^p) \le y/(y^p-(r/2)^p) = g(y)$ (because $y - x \le y$ and $y^p - x^p \ge y^p - (r/2)^p$), we obtain $|x-y| \le r^{1-p}/(1-2^{-p}) \cdot |x^p - y^p|$. For $p = 1$: $|x-y| \le |x-y|$ trivially, and $C_1 := 1$ suffices.

In both cases $|x-y| \le C_p\, r^{1-p}|x^p - y^p|$ with $C_p := \max(2^{p-1}/p,\ 1/(1-2^{-p}))$. □

Theorem 10.3 (Fixed-annulus completion criterion)**.** *Assume $p \ge 2$, $d > 2p$, and $t \in [\eta_N, 1]$. If an estimator $\widehat{T}_t$ satisfies $\sup_{(P,Q)\in\mathcal{A}_t} \mathbb{E}|\widehat{T}_t - T^{\mathrm{far}}_{t,p}(P,Q)| \le C_1\, \eta_N t^{p-1}$ and squared error $\le C_2\, \eta_N^2 t^{2p-2}$, then the distance-level estimator achieves risk $O(\eta_N)$ on $\mathcal{A}_t$.*

Proof. Define $\widehat{W}_t := (\widehat{T}_t)^{1/p}$ (clipped to $[0, \sqrt{d}]$). On $\mathcal{A}_t$: $W_p(P,Q) \ge t/2$, so by Theorem 10.2 with $r = t/2$: $|\widehat{W}_t - W_p(P,Q)| \le C_p'(t/2)^{1-p}|\widehat{T}_t - W_p^p| \le C_p' t^{1-p}(|\widehat{T}_t - T^{\mathrm{far}}| + |T^{\mathrm{far}} - W_p^p|)$. By the adaptive linearization (Theorem 9.1), $|T^{\mathrm{far}} - W_p^p| \le C\eta_N t^{p-1}$. Taking expectations: $\mathbb{E}|\widehat{W}_t - W_p| \le C_p t^{1-p}(C_1\eta_N t^{p-1} + C\eta_N t^{p-1}) = O(\eta_N)$. □

11 Shell-active semiconcavity and the failure of the polyhedral continuation

The annular localization from the previous sections isolates the genuinely difficult residual regime. A natural next hope would be that, after truncation to the far-offset functional, optimal dual potentials become cellwise polyhedral on each dyadic cube. This is false. The correct surviving structure is *shell-active* rather than polyhedral.

Throughout this section we write

$$c_p(x,y) := \|x-y\|^p .$$

Proposition 11.1 (The cellwise polyhedral continuation is false)**.** *Assume $p \ge 2$. There exist a bounded Borel function $\psi : [0,1]^d \to \mathbb{R}$ and a dyadic cell $R \subset [0,1]^d$ such that the c_p-transform $\psi^{c_p}|_R$ is not polyhedral concave. In fact, it is not even concave on R.*

Proof. Fix $y_0 \in (0,1)^d$ and a dyadic cell R whose closure does not contain y_0. Let

$$M > 2\ \mathrm{diam}([0,1]^d)^p = 2d^{p/2},$$

and define

$$\psi(y_0) := 0, \qquad \psi(y) := -M \quad (y \ne y_0).$$

Then for every $x \in [0,1]^d$,

$$\psi^{c_p}(x) = \inf_{y\in[0,1]^d} \big(\|x-y\|^p - \psi(y)\big) = \min\Big\{ \|x-y_0\|^p,\ \inf_{y\neq y_0} \big(\|x-y\|^p + M\big)\Big\}.$$

Because $\|x-y\|^p \le d^{p/2}$ on $[0,1]^d \times [0,1]^d$, the second term is at least $M > d^{p/2}$, whereas the first is at most $d^{p/2}$. Hence

$$\psi^{c_p}(x) = \|x-y_0\|^p \qquad \forall x \in [0,1]^d.$$

For $p \ge 2$ the map $x \mapsto \|x-y_0\|^p$ is convex, and on any cell R disjoint from y_0 it is not concave. Therefore $\psi^{c_p}|_R$ cannot be polyhedral concave. □

We now replace the false polyhedral picture by a correct shell-active description of optimal dual potentials.

11.1 Nearest active branches

Let (ϕ, ψ) be an optimal dual pair for $W_p(P,Q)^p$:

$$\phi(x) + \psi(y) \le c_p(x,y) \quad \forall x, y, \qquad \int \phi\, dP + \int \psi\, dQ = W_p(P,Q)^p.$$

Since the cost is continuous on a compact set, we may and do assume that ϕ and ψ are continuous c_p-transforms of each other. Define the active set

$$\Gamma_\psi(y) := \operatorname{argmin}_{x\in[0,1]^d} \big(c_p(x,y) - \phi(x)\big).$$

This set is nonempty and compact for every y.

Lemma 11.2 (Measurable nearest active branch)**.** *There exists a Borel measurable map*

$$x_\star : [0,1]^d \to [0,1]^d$$

such that for every $y \in [0,1]^d$,

$$x_\star(y) \in \Gamma_\psi(y), \qquad \|x_\star(y) - y\| = \min_{x\in\Gamma_\psi(y)} \|x-y\|\,.$$

If π is any optimal coupling between P and Q, then

$$\int \|x_\star(y) - y\|^p \, dQ(y) \le \int \|x-y\|^p \, d\pi(x,y) = W_p(P,Q)^p.$$

Proof. The map

$$F(y,x) := c_p(x,y) - \phi(x)$$

is continuous on the compact set $[0,1]^d \times [0,1]^d$. Hence $y \mapsto \Gamma_\psi(y) = \operatorname{argmin}_x F(y,x)$ is a measurable compact-valued multifunction by the measurable maximum theorem. The secondary objective $x \mapsto \|x-y\|$ is continuous, so the multifunction

$$\widetilde{\Gamma}_\psi(y) := \operatorname{argmin}_{x\in\Gamma_\psi(y)} \|x-y\|$$

is again measurable, compact-valued, and nonempty. A Kuratowski–Ryll-Nardzewski selection yields a Borel measurable selector $x_\star$.

Now let π be an optimal coupling. By complementary slackness,

$$\phi(x) + \psi(y) = c_p(x, y) \qquad \pi\text{-a.e.}$$

Since the equality set

$$\{(x, y) : \phi(x) + \psi(y) = c_p(x, y)\}$$

is closed and has full π-measure, it contains $\operatorname{supp} \pi$. Therefore, for Q-a.e. y, every point of $\operatorname{supp}(\pi_y)$ belongs to $\Gamma_\psi(y)$, where $\pi(dx, dy) = \pi_y(dx)\, Q(dy)$ is a disintegration of π. Hence

$$\|x_\star(y) - y\|^p \le \inf_{x \in \operatorname{supp}(\pi_y)} \|x - y\|^p \le \int \|x - y\|^p \, d\pi_y(x)$$

for Q-a.e. y. Integrating in y yields the claim. □

Definition 11.3 (Nearest active radius)**.** For a measurable nearest active branch $x_\star$ as in Theorem 11.2, define

$$r_\psi(y) := \|x_\star(y) - y\| .$$

11.2 Local semiconcavity at the active scale

Proposition 11.4 (Active-branch local quadratic upper bound)**.** *Assume $p \ge 2$. Let (ϕ, ψ) be an optimal dual pair and let $x_\star$ be a measurable nearest active branch. Then for every $y \in [0, 1]^d$ and every $z \in [0, 1]^d$ satisfying*

$$\|z - y\| \le \frac{r_\psi(y)}{2},$$

one has

$$\psi(z) \le \psi(y) + \nabla_y c_p(x_\star(y), y) \cdot (z - y) + L_p\, r_\psi(y)^{p-2} \|z - y\|^2 ,$$

where

$$L_p := \frac{p(p-1)}{2} \Big(\frac{3}{2}\Big)^{p-2}.$$

Equivalently, whenever $\|h\| \le r_\psi(y)/2$,

$$\psi(y + h) + \psi(y - h) - 2\psi(y) \le 2L_p\, r_\psi(y)^{p-2} \|h\|^2 .$$

Proof. Fix y and write $x := x_\star(y)$, $r := r_\psi(y) = \|x - y\|$. Since $x \in \Gamma_\psi(y)$,

$$\psi(y) = c_p(x, y) - \phi(x).$$

For any z,

$$\psi(z) \le c_p(x, z) - \phi(x) = \psi(y) + \big(c_p(x, z) - c_p(x, y)\big).$$

Define

$$f_x(u) := c_p(x, u) = \|x - u\|^p .$$

Then f_x is C^2 for $p \ge 2$, and

$$\nabla^2 f_x(u) = p\, \|x - u\|^{p-2} I + p(p-2)\, \|x - u\|^{p-4} (x - u)(x - u)^\top .$$

Hence

$$\big\|\nabla^2 f_x(u)\big\|_{\mathrm{op}} \le p(p-1)\, \|x - u\|^{p-2} .$$

If $\|z-y\| \le r/2$, then every point u on the segment $[y,z]$ satisfies

$$\|x-u\| \le \|x-y\| + \|u-y\| \le r + \frac{r}{2} = \frac{3r}{2}.$$

Therefore

$$\sup_{u\in[y,z]} \left\|\nabla^2 f_x(u)\right\|_{\mathrm{op}} \le p(p-1)\Big(\frac{3r}{2}\Big)^{p-2}.$$

Taylor's theorem gives

$$f_x(z) \le f_x(y) + \nabla f_x(y)\cdot(z-y) + \frac{p(p-1)}{2}\Big(\frac{3r}{2}\Big)^{p-2}\|z-y\|^2.$$

Combining with $\psi(z)-\psi(y) \le f_x(z)-f_x(y)$ proves the first inequality. The second follows by applying the first bound to $z = y+h$ and $z = y-h$ and summing. □

Corollary 11.5 (Shell decomposition with quantitative masses)**.** *Assume $p \ge 2$ and let (ϕ,ψ) be an optimal dual pair for (P,Q). Let $x_\star$ and r_ψ be as above. For $t>0$, define*

$$A_{-1}(t) := \{y : r_\psi(y) < t\}, \qquad A_j(t) := \{y : 2^j t \le r_\psi(y) < 2^{j+1}t\} \quad (j \ge 0).$$

Then:

(i) *For every $j \ge 0$,*

$$Q(A_j(t)) \le 2^{-jp}\, t^{-p}\, W_p(P,Q)^p.$$

In particular, on the annulus $W_p(P,Q) \in [t/2, 2t]$,

$$Q(A_j(t)) \le 2^p\, 2^{-jp}.$$

(ii) *For every $j \ge 0$, every $y \in A_j(t)$, and every z with*

$$\|z-y\| \le 2^{j-1}t,$$

one has

$$\psi(z) \le \psi(y) + \nabla_y c_p(x_\star(y), y)\cdot(z-y) + L_p\,(2^{j+1}t)^{p-2}\|z-y\|^2.$$

(iii) *For every $j \ge 0$ and every $y \in A_j(t)$,*

$$\|\nabla_y c_p(x_\star(y), y)\| = p\, r_\psi(y)^{p-1} \le p\,(2^{j+1}t)^{p-1}.$$

Proof. Part (i) is Markov's inequality together with Theorem 11.2:

$$Q(A_j(t)) \le \frac{1}{(2^j t)^p}\int r_\psi(y)^p\, dQ(y) \le \frac{W_p(P,Q)^p}{(2^j t)^p}.$$

Parts (ii) and (iii) follow immediately from Theorem 11.4 and the identity

$$\nabla_y c_p(x,y) = p\,\|x-y\|^{p-2}\,(y-x).$$ □

Remark 11.6. Theorem 11.5 replaces the false polyhedral picture with a correct structural description. What survives is not a finite-alphabet representation of the dual class, but a

shell decomposition in which the curvature scale and the shell mass are both controlled by the active radius $r_\psi(y)$.

11.3 Deterministic-support sparse control

For the next proposition we write

$$\Psi_p^{c,0} := \{\psi : [0,1]^d \to \mathbb{R} : \ \psi \text{ is continuous and } c_p\text{-concave}, \ \psi(0) = 0\}.$$

This is the same normalized class that will be studied globally in Section 14.

Lemma 11.7 (Binomial thinning)**.** *Let Q be a probability measure on $[0,1]^d$, let $A \subset [0,1]^d$ be measurable with*

$$w := Q(A) \in [0,1],$$

and let $\mathcal{G}$ be any class of bounded measurable functions on $[0,1]^d$. Assume that for each $M \geq 1$ and each probability measure μ on A there exists a deterministic bound Γ_M such that

$$\mathbb{E} \sup_{g \in \mathcal{G}} |(\mu_M - \mu)g| \leq \Gamma_M,$$

where μ_M is the empirical measure of M i.i.d. samples from μ. Let

$$B := \sup_{g \in \mathcal{G}} \sup_{x \in [0,1]^d} |g(x)|.$$

Then, for i.i.d. $Y_1, \dots, Y_N \sim Q$,

$$\mathbb{E} \sup_{g \in \mathcal{G}} |(Q_N - Q)(g\mathbf{1}_A)| \leq \mathbb{E}\Big[\frac{M}{N}\Gamma_M \, \mathbf{1}_{\{M \geq 1\}}\Big] + B\sqrt{\frac{w}{N}},$$

where

$$M := \sum_{i=1}^N \mathbf{1}_A(Y_i) \sim \mathrm{Bin}(N, w).$$

Proof. If $w = 0$, then $M = 0$ almost surely and the left-hand side is 0, so there is nothing to prove. Assume henceforth that $w > 0$ and let

$$Q_A := Q(\cdot \mid A).$$

Condition on M and on the event $M = m$. If $m = 0$, then $Q_N(g\mathbf{1}_A) = 0$ for every g and the centered process is simply $-Q(g\mathbf{1}_A)$, whose contribution is bounded by

$$B\, Q(A) = Bw = B\Big|\frac{m}{N} - w\Big|.$$

If $m \geq 1$, let $Q_{A,m}$ denote the empirical measure of the m sample points falling in A. Then for every $g \in \mathcal{G}$,

$$Q_N(g\mathbf{1}_A) - Q(g\mathbf{1}_A) = \frac{m}{N}(Q_{A,m} - Q_A)g + \Big(\frac{m}{N} - w\Big)Q_A g.$$

Hence

$$\sup_{g \in \mathcal{G}} |(Q_N - Q)(g\mathbf{1}_A)| \leq \frac{m}{N} \sup_{g \in \mathcal{G}} |(Q_{A,m} - Q_A)g| + B\Big|\frac{m}{N} - w\Big|.$$

Taking conditional expectations and then averaging in M yields

$$\mathbb{E}\sup_{g\in\mathcal{G}}|(Q_N-Q)(g\mathbf{1}_A)| \le \mathbb{E}\Big[\frac{M}{N}\Gamma_M\,\mathbf{1}_{\{M\ge 1\}}\Big] + B\,\mathbb{E}\Big|\frac{M}{N}-w\Big|.$$

Since $M\sim \mathrm{Bin}(N,w)$,

$$\mathbb{E}\Big|\frac{M}{N}-w\Big| \le \sqrt{\mathrm{Var}(M/N)} = \sqrt{\frac{w(1-w)}{N}} \le \sqrt{\frac{w}{N}}. \qquad \square$$

Proposition 11.8 (Fixed-support sparse semiconcavity)**.** *Assume $d\ge 6$ and $p\ge 2$. There exists a constant $C_{d,p}$ such that for every probability measure Q on $[0,1]^d$, every measurable set $A\subset[0,1]^d$ with $Q(A)=w$, and every $N\ge 2$,*

$$\mathbb{E}\sup_{\psi\in\Psi_p^{c,0}}|(Q_N-Q)(\psi\,\mathbf{1}_A)| \le C_{d,p}\Big(N^{-2/(d-1)}w^{1-2/(d-1)} + \sqrt{\frac{w}{N}}\Big).$$

Proof. Apply Theorem 11.7 with $\mathcal{G}=\Psi_p^{c,0}$ and

$$\Gamma_M := C_{d,p}M^{-2/(d-1)},$$

which is available from the global bound Theorem 16.1 for the normalized class. Since c_p is bounded on $[0,1]^d\times[0,1]^d$ and $\psi(0)=0$, every $\psi\in\Psi_p^{c,0}$ satisfies

$$|\psi(x)| \le \mathrm{diam}([0,1]^d)^p = d^{p/2} \qquad \forall x\in[0,1]^d,$$

so the envelope constant B is bounded by $d^{p/2}$. Therefore,

$$\mathbb{E}\sup_{\psi\in\Psi_p^{c,0}}|(Q_N-Q)(\psi\,\mathbf{1}_A)| \le C_{d,p}\,\mathbb{E}\Big[\frac{M}{N}M^{-2/(d-1)}\mathbf{1}_{\{M\ge 1\}}\Big] + d^{p/2}\sqrt{\frac{w}{N}}.$$

Since $d\ge 6$, the exponent

$$\alpha := 1-\frac{2}{d-1}$$

lies in $(0,1)$, and $x\mapsto x^\alpha$ is concave on $[0,\infty)$. Hence

$$\mathbb{E}\Big[\frac{M}{N}M^{-2/(d-1)}\mathbf{1}_{\{M\ge 1\}}\Big] \le \frac{1}{N}\mathbb{E}[M^\alpha] \le \frac{(\mathbb{E}M)^\alpha}{N} = N^{-2/(d-1)}w^{1-2/(d-1)}.$$

Absorbing the envelope term into the constant proves the claim. $\square$

Remark 11.9. Theorem 11.8 shows that, once the support is deterministic, the shell mass w automatically enters with exactly the exponent needed to sum the shell decomposition from Theorem 11.5. The real obstruction is therefore not the shell mass itself, but the fact that the shell support depends on the unknown optimal potential.

12 A fixed-potential transport inequality on annuli

Even though the adaptive-shell transfer principle fails, one robust piece of structure survives: a *fixed* optimal dual potential can still be transported against an arbitrary perturbation of the second marginal at exactly the annular homogeneity dictated by its active radii.

Proposition 12.1 (Global one-sided increment bound)**.** *Assume $p \geq 2$. Let (ϕ, ψ) be an optimal dual pair and let $x_\star$ be a measurable nearest active branch with active radius r_ψ. Then for all $y, z \in [0,1]^d$,*

$$\psi(z) - \psi(y) \leq p\, r_\psi(y)^{p-1} \|z - y\| + L_p\, r_\psi(y)^{p-2} \|z - y\|^2 + 3^p \|z - y\|^p .$$

Proof. Fix y, z and write

$$h := z - y, \qquad r := r_\psi(y), \qquad x := x_\star(y).$$

If $\|h\| \leq r/2$, then Theorem 11.4 gives

$$\psi(z) - \psi(y) \leq \nabla_y c_p(x, y) \cdot h + L_p r^{p-2} \|h\|^2 .$$

Since

$$\|\nabla_y c_p(x, y)\| = p\, r^{p-1},$$

we obtain

$$\psi(z) - \psi(y) \leq p\, r^{p-1} \|h\| + L_p r^{p-2} \|h\|^2 .$$

If $\|h\| > r/2$, then $r < 2\|h\|$ and

$$\psi(z) - \psi(y) \leq c_p(x, z) - c_p(x, y) \leq \|x - z\|^p \leq (r + \|h\|)^p \leq 3^p \|h\|^p .$$

Combining the two cases proves the claim. □

Theorem 12.2 (Deterministic fixed-potential transport control)**.** *Assume $p \geq 2$, and for $p = 2$ interpret $a^{(p-2)/p}$ as 1 for all $a \geq 0$. Let (ϕ, ψ) be an optimal dual pair for (P, Q), let r_ψ be the active radius of a measurable nearest active branch, and let $\mu, \nu \in \mathcal{P}([0,1]^d)$ be arbitrary. Then*

$$\begin{aligned} \left|\int \psi\, d\mu - \int \psi\, d\nu\right| \leq{}& p\Big(\Big(\int r_\psi^p\, d\mu\Big)^{\frac{p-1}{p}} + \Big(\int r_\psi^p\, d\nu\Big)^{\frac{p-1}{p}}\Big) W_p(\mu, \nu) \\ &+ L_p\Big(\Big(\int r_\psi^p\, d\mu\Big)^{\frac{p-2}{p}} + \Big(\int r_\psi^p\, d\nu\Big)^{\frac{p-2}{p}}\Big) W_p(\mu, \nu)^2 + 2 \cdot 3^p\, W_p(\mu, \nu)^p. \end{aligned}$$

Proof. Let $\gamma \in \Pi(\mu, \nu)$ be an optimal coupling, so that

$$\int \|z - y\|^p \, d\gamma(y, z) = W_p(\mu, \nu)^p.$$

Then

$$\int \psi\, d\nu - \int \psi\, d\mu = \int (\psi(z) - \psi(y))\, d\gamma(y, z).$$

Applying Theorem 12.1 yields

$$\begin{aligned} \int \psi\, d\nu - \int \psi\, d\mu \leq{}& p \int r_\psi(y)^{p-1} \|z - y\| \, d\gamma(y, z) \\ &+ L_p \int r_\psi(y)^{p-2} \|z - y\|^2 \, d\gamma(y, z) + 3^p \int \|z - y\|^p \, d\gamma(y, z). \end{aligned}$$

For the first term, Hölder's inequality with exponents $p/(p-1)$ and p gives

$$\int r_\psi(y)^{p-1} \|z - y\| \, d\gamma(y, z) \leq \Big(\int r_\psi(y)^p \, d\gamma(y, z)\Big)^{\frac{p-1}{p}} \Big(\int \|z - y\|^p \, d\gamma(y, z)\Big)^{\frac{1}{p}}.$$

Since the first marginal of γ is μ,

$$\int r_\psi(y)^p \, d\gamma(y,z) = \int r_\psi^p \, d\mu.$$

Hence

$$\int r_\psi(y)^{p-1} \, \|z-y\| \, d\gamma(y,z) \le \Big(\int r_\psi^p \, d\mu\Big)^{\frac{p-1}{p}} W_p(\mu,\nu).$$

For the second term, if $p > 2$ then Hölder with exponents $p/(p-2)$ and $p/2$ gives

$$\int r_\psi(y)^{p-2} \, \|z-y\|^2 \, d\gamma(y,z) \le \Big(\int r_\psi^p \, d\mu\Big)^{\frac{p-2}{p}} W_p(\mu,\nu)^2.$$

When $p = 2$, the same display reduces to the identity

$$\int \|z-y\|^2 \, d\gamma(y,z) = W_2(\mu,\nu)^2.$$

The last term is simply $3^p W_p(\mu,\nu)^p$. Therefore

$$\begin{aligned}\int \psi \, d\nu - \int \psi \, d\mu \le{}& p\Big(\int r_\psi^p \, d\mu\Big)^{\frac{p-1}{p}} W_p(\mu,\nu) \\ &+ L_p\Big(\int r_\psi^p \, d\mu\Big)^{\frac{p-2}{p}} W_p(\mu,\nu)^2 + 3^p W_p(\mu,\nu)^p.\end{aligned}$$

Interchanging μ and ν gives the analogous bound for $\int \psi \, d\mu - \int \psi \, d\nu$. Adding the two one-sided bounds implies the stated estimate for the absolute value. □

Corollary 12.3 (Empirical fluctuation of a fixed annular optimal potential)**.** *Assume $p \ge 2$, let (ϕ,ψ) be an optimal dual pair for (P,Q), and let*

$$\epsilon_{N,p}(Q) := \Big(\mathbb{E} W_p(Q_N,Q)^p\Big)^{1/p},$$

where Q_N is the empirical measure of N i.i.d. samples from Q. Then

$$\mathbb{E}\big|(Q_N - Q)\psi\big| \le 2p\Big(\int r_\psi^p \, dQ\Big)^{\frac{p-1}{p}} \epsilon_{N,p}(Q) + 2L_p\Big(\int r_\psi^p \, dQ\Big)^{\frac{p-2}{p}} \epsilon_{N,p}(Q)^2 + 2\cdot 3^p \, \epsilon_{N,p}(Q)^p.$$

Consequently, if $d > 2p$, then

$$\mathbb{E}\big|(Q_N - Q)\psi\big| \lesssim_{d,p} W_p(P,Q)^{p-1} N^{-1/d} + W_p(P,Q)^{p-2} N^{-2/d} + N^{-p/d}.$$

In particular, on the annulus $W_p(P,Q) \in [t/2, 2t]$,

$$\mathbb{E}\big|(Q_N - Q)\psi\big| \lesssim_{d,p} t^{p-1} N^{-1/d} + t^{p-2} N^{-2/d} + N^{-p/d}.$$

Proof. Apply Theorem 12.2 with $\mu = Q$ and $\nu = Q_N$, then take expectations. The terms involving $\int r_\psi^p \, dQ_N$ are controlled by Hölder's inequality:

$$\mathbb{E}\Big[\Big(\int r_\psi^p \, dQ_N\Big)^{\frac{p-1}{p}} W_p(Q_N,Q)\Big] \le \Big(\mathbb{E}\int r_\psi^p \, dQ_N\Big)^{\frac{p-1}{p}} \Big(\mathbb{E} W_p(Q_N,Q)^p\Big)^{\frac{1}{p}},$$

and, for $p > 2$,

$$\mathbb{E}\Big[\Big(\int r_\psi^p \, dQ_N\Big)^{\frac{p-2}{p}} W_p(Q_N, Q)^2\Big] \le \Big(\mathbb{E}\int r_\psi^p \, dQ_N\Big)^{\frac{p-2}{p}} \Big(\mathbb{E}W_p(Q_N, Q)^p\Big)^{\frac{2}{p}}.$$

When $p = 2$, the corresponding term is simply $\mathbb{E}W_2(Q_N, Q)^2$. Since

$$\mathbb{E}\int r_\psi^p \, dQ_N = \int r_\psi^p \, dQ$$

and Theorem 11.2 gives

$$\int r_\psi^p \, dQ \le W_p(P, Q)^p,$$

the first estimate follows.

For the second estimate, use the standard supercritical one-sample bound

$$\epsilon_{N,p}(Q) \lesssim_{d,p} N^{-1/d}$$

uniformly over $Q \in \mathcal{P}([0,1]^d)$; see, for example, [9, 18]. This yields the displayed rate bound, and the annular form follows by substituting $W_p(P, Q) \asymp t$. □

Remark 12.4 (A logarithmic fixed-potential barrier). At the critical annulus $t = \eta_N$, Theorem 12.3 gives the cost-level scale

$$N^{-p/d}(\log N)^{-(p-1)/d} + N^{-p/d}(\log N)^{-(p-2)/d} + N^{-p/d},$$

whereas the minimax target is

$$\eta_N^p = N^{-p/d}(\log N)^{-p/d}.$$

Thus a direct argument based on a *single* population optimal potential and its transport under one empirical perturbation reaches the correct homogeneity in t but still misses the residual superquadratic target by logarithmic factors. Any complete closure beyond the semiconcave wedge must therefore exploit a genuinely different cancellation mechanism.

13 A no-go theorem for adaptive shell transfer

The previous section suggests an ambitious continuation strategy: combine the shell mass estimate from Theorem 11.5 with the deterministic-support sparse bound from Theorem 11.8, then sum over shells. The next theorem shows that this adaptive step is impossible in general.

Proposition 13.1 (Shell classes shatter separated configurations)**.** *Let $d \ge 1$, $p \ge 1$, and $t \in (0, 1/24]$. Let*

$$y_1, \dots, y_M \in [1/3, 2/3]^d$$

satisfy

$$\|y_i - y_j\| \ge 8t \qquad (i \ne j).$$

Then for every subset $S \subset [M]$ there exist a probability measure

$$Q := \frac{1}{M}\sum_{i=1}^{M} \delta_{y_i},$$

a probability measure

$$P_S := \frac{1}{M}\sum_{i=1}^{M} \delta_{x_i^S},$$

and a normalized optimal dual potential $\psi_S \in \Psi_p^{c,0}$ *for* (P_S, Q) *such that the following hold.*

(i) *Writing*

$$r_i^S := \begin{cases} \dfrac{3t}{2}, & i \in S, \\ \dfrac{t}{2}, & i \notin S, \end{cases} \qquad x_i^S := y_i + r_i^S e_1,$$

one has

$$W_p(P_S, Q)^p = \frac{1}{M}\sum_{i=1}^{M} (r_i^S)^p.$$

(ii) *If* $x_\star$ *is the measurable nearest active branch from Theorem 11.2 applied to* ψ_S*, then*

$$r_{\psi_S}(y_i) = r_i^S \qquad (1 \le i \le M).$$

Consequently,

$$A_0^{\psi_S}(t) \cap \{y_1, \dots, y_M\} = \{y_i : \ i \in S\},$$

where

$$A_0^{\psi_S}(t) := \{y : \ t \le r_{\psi_S}(y) < 2t\}.$$

(iii) *There exists a constant*

$$c_p := 3^{-p} - 16^{-p} > 0$$

such that

$$\psi_S(y_i) \le -c_p \qquad \forall i = 1, \dots, M.$$

Proof. For the given subset S, define r_i^S and x_i^S as in the statement and set

$$\psi_{S,0}(y) := \min_{1 \le i \le M} \left(\left\| x_i^S - y \right\|^p - (r_i^S)^p \right).$$

This is a continuous c_p-concave function as a finite minimum of c_p-affine functions. Normalize it by

$$\psi_S(y) := \psi_{S,0}(y) - \psi_{S,0}(0),$$

so that $\psi_S(0) = 0$.

Fix i. The i-th branch gives

$$\left\| x_i^S - y_i \right\|^p - (r_i^S)^p = 0.$$

If $j \ne i$, then by the separation assumption,

$$\left\| x_j^S - y_i \right\| \ge \|y_j - y_i\| - r_j^S \ge 8t - \frac{3t}{2} = \frac{13t}{2},$$

hence

$$\left\| x_j^S - y_i \right\|^p - (r_j^S)^p \ge \left(\frac{13t}{2}\right)^p - \left(\frac{3t}{2}\right)^p > 0.$$

Therefore the minimum defining $\psi_{S,0}(y_i)$ is attained uniquely at index i, and

$$\psi_{S,0}(y_i) = 0 \qquad \forall i.$$

In particular,

$$\psi_S(y_i) = -\psi_{S,0}(0) \qquad \forall i.$$

Since each x_i^S lies in $[1/3, 3/4] \times [1/3, 2/3]^{d-1}$ and $r_i^S \le 3t/2 \le 1/16$, we obtain

$$\psi_{S,0}(0) = \min_i \big(\big\| x_i^S \big\|^p - (r_i^S)^p \big) \ge \Big(\frac{1}{3}\Big)^p - \Big(\frac{1}{16}\Big)^p = c_p > 0.$$

This proves part (iii).

Now let $\phi_S := \psi_S^{c_p}$. Because

$$\psi_{S,0}(y) \le \big\| x_i^S - y \big\|^p - (r_i^S)^p \qquad \forall y,$$

we have

$$\phi_S(x_i^S) = \inf_y \big(\big\| x_i^S - y \big\|^p - \psi_S(y) \big) = \psi_{S,0}(0) + \inf_y \big(\big\| x_i^S - y \big\|^p - \psi_{S,0}(y) \big) \ge \psi_{S,0}(0) + (r_i^S)^p.$$

At $y = y_i$ equality holds because $\psi_{S,0}(y_i) = 0$. Hence

$$\phi_S(x_i^S) = \psi_{S,0}(0) + (r_i^S)^p \qquad \forall i.$$

Therefore

$$\frac{1}{M} \sum_{i=1}^M \phi_S(x_i^S) + \frac{1}{M} \sum_{i=1}^M \psi_S(y_i) = \frac{1}{M} \sum_{i=1}^M (r_i^S)^p.$$

On the other hand, the coupling that matches x_i^S to y_i has exactly the same cost. By Kantorovich duality this coupling is optimal and

$$W_p(P_S, Q)^p = \frac{1}{M} \sum_{i=1}^M (r_i^S)^p,$$

which proves part (i).

Finally, at each support point y_i the unique active branch is x_i^S. The nearest active radius is therefore

$$r_{\psi_S}(y_i) = \big\| x_i^S - y_i \big\| = r_i^S.$$

Hence $y_i \in A_0^{\psi_S}(t)$ exactly when $r_i^S = 3t/2$, that is, exactly when $i \in S$. This proves part (ii). □

Corollary 13.2 (Infinite VC dimension of shell classes)**.** *For every $p \ge 1$, the family*

$$\{ A_0^{\psi}(t) : \ t \in (0, 1/24], \ \psi \in \Psi_p^{c,0} \}$$

has infinite VC dimension, even when ψ ranges only over normalized optimal dual potentials of pairs (P, Q) with $W_p(P, Q) \asymp t$.

Proof. For any finite separated set $\{y_1, \dots, y_M\}$ as in Theorem 13.1, every subset $S \subset [M]$ is realized as

$$A_0^{\psi_S}(t) \cap \{y_1, \dots, y_M\}.$$

Since M is arbitrary, the VC dimension is infinite. □

Lemma 13.3 (Half-subsets capture total variation)**.** *Let $M = 2k$ be even and let $d_1, \dots, d_M \in \mathbb{R}$ satisfy*

$$\sum_{i=1}^{M} d_i = 0.$$

Then

$$\sup_{\substack{S\subset[M]\\|S|=k}} \sum_{i\in S} d_i \geq \frac{1}{2}\sum_{i=1}^{M} d_i^+ = \frac{1}{4}\sum_{i=1}^{M} |d_i|.$$

Proof. Let $P := \{i : \ d_i > 0\}$ and write $r := |P|$ and

$$V := \sum_{i=1}^{M} d_i^+ = \frac{1}{2}\sum_{i=1}^{M} |d_i|.$$

If $r > k$, choose S to be the indices of the k largest positive entries. Then

$$\sum_{i\in S} d_i \geq \frac{k}{r}\sum_{i\in P} d_i = \frac{k}{r}V \geq \frac{V}{2},$$

because $r \leq 2k$.

If $r \leq k$, choose S by taking all positive indices together with the $k - r$ largest nonpositive entries. Let

$$B := \sum_{i\in S\setminus P} |d_i|.$$

Because the selected nonpositive entries are the least negative ones, their average absolute value is at most the average absolute value of the remaining nonpositive entries. The latter set has cardinality

$$(2k - r) - (k - r) = k.$$

Hence

$$\frac{B}{k-r} \leq \frac{V-B}{k},$$

which implies

$$B \leq \frac{k-r}{2k-r}\, V.$$

Therefore

$$\sum_{i\in S} d_i = V - B \geq V\Big(1 - \frac{k-r}{2k-r}\Big) = \frac{k}{2k-r}V \geq \frac{V}{2}.$$

This proves the claim. □

Theorem 13.4 (Quantitative failure of adaptive shell transfer)**.** *Assume $d \geq 6$ and $p \geq 2$. There exist constants $\kappa_d \in (0, 1/48)$, $C_p \geq 1$, and $c_{d,p} > 0$ such that for every even N large enough one can find a probability measure $\overline{Q}_N \in \mathcal{P}([0,1]^d)$ and a scale*

$$t_N := \kappa_d N^{-1/d} \in [\eta_N, 1/24]$$

with the following property.

If $\overline{Q}_{N,\mathrm{emp}}$ denotes the empirical measure of N i.i.d. samples from $\overline{Q}_N$, then there exists a family of normalized optimal dual potentials

$$\{\psi_S : \ S \subset [N],\ |S| = N/2\} \subset \Psi_p^{c,0}$$

and associated measures $P_{N,S}$ such that

$$C_p^{-1} t_N \le W_p(P_{N,S}, \overline{Q}_N) \le C_p t_N \qquad \forall S,\ |S| = N/2,$$

and

$$\mathbb{E} \sup_{\substack{S\subset[N]\\ |S|=N/2}} \left|(\overline{Q}_{N,\mathrm{emp}} - \overline{Q}_N)(\psi_S \mathbf{1}_{A_0^{\psi_S}(t_N)})\right| \ge c_{d,p}.$$

Consequently, no uniform inequality of adaptive-shell-transfer type can hold with a right-hand side tending to zero. More precisely, there is no sequence $\varepsilon_N \to 0$ such that

$$\mathbb{E}\sup\Big\{ \left|(Q_N - Q)(\psi \mathbf{1}_{A_0^{\psi}(t)})\right| : \begin{array}{l} \psi \in \Psi_p^{c,0} \text{ is an optimal dual potential for some pair } (P,Q), \\ W_p(P,Q) \in [t/2, 2t] \end{array} \Big\} \le \varepsilon_N$$

uniformly over all $Q \in \mathcal{P}([0,1]^d)$, all $t \in [\eta_N, 1]$, and all N.

Proof. Fix $\kappa_d > 0$ so small that for every large N the cube $[1/3, 2/3]^d$ contains at least N points with pairwise separation at least $8\kappa_d N^{-1/d}$. For instance, one may take a regular grid of mesh $8\kappa_d N^{-1/d}$ and choose κ_d so that the number of grid points is at least N. Set

$$t_N := \kappa_d N^{-1/d}.$$

Then $t_N \le 1/24$ for all large N, and

$$\frac{t_N}{\eta_N} = \kappa_d (\log N)^{1/d} \to \infty,$$

so $t_N \ge \eta_N$ eventually.

Choose such a separated family

$$y_1^{(N)}, \dots, y_N^{(N)} \in [1/3, 2/3]^d$$

and define

$$\overline{Q}_N := \frac{1}{N} \sum_{i=1}^N \delta_{y_i^{(N)}}.$$

For each subset $S \subset [N]$ with $|S| = N/2$, apply Theorem 13.1 with $M = N$, $t = t_N$, and points $y_i = y_i^{(N)}$. This yields a measure $P_{N,S}$ and a normalized optimal dual potential ψ_S satisfying

$$W_p(P_{N,S}, \overline{Q}_N)^p = \frac{1}{2}\Big(\frac{3t_N}{2}\Big)^p + \frac{1}{2}\Big(\frac{t_N}{2}\Big)^p = 2^{-p-1}(3^p + 1) t_N^p.$$

Hence

$$C_p^{-1} t_N \le W_p(P_{N,S}, \overline{Q}_N) \le C_p t_N$$

for a constant C_p depending only on p.

Let $\overline{Q}_{N,\mathrm{emp}}$ be the empirical measure of N i.i.d. samples from $\overline{Q}_N$. By Theorem 13.1(ii)–(iii), each function

$$f_S := \psi_S \mathbf{1}_{A_0^{\psi_S}(t_N)}$$

satisfies

$$f_S(y_i^{(N)}) = \begin{cases} -c_S, & i \in S, \\ 0, & i \notin S, \end{cases} \qquad c_S \geq c_p := 3^{-p} - 16^{-p} > 0.$$

Therefore

$$|(\overline{Q}_{N,\mathrm{emp}} - \overline{Q}_N) f_S| = c_S \Big|\overline{Q}_{N,\mathrm{emp}}(S) - \frac{1}{2}\Big| \geq c_p \Big|\overline{Q}_{N,\mathrm{emp}}(S) - \frac{1}{2}\Big|.$$

Taking the supremum over all $|S| = N/2$ gives

$$\sup_{|S|=N/2} |(\overline{Q}_{N,\mathrm{emp}} - \overline{Q}_N) f_S| \geq c_p \sup_{|S|=N/2} \Big|\overline{Q}_{N,\mathrm{emp}}(S) - \frac{1}{2}\Big|.$$

Write

$$K_i := N\, \overline{Q}_{N,\mathrm{emp}}(\{y_i^{(N)}\}), \qquad d_i := \frac{K_i}{N} - \frac{1}{N}.$$

Then $\sum_i d_i = 0$ and, since N is even, Theorem 13.3 yields

$$\sup_{|S|=N/2} \Big(\overline{Q}_{N,\mathrm{emp}}(S) - \frac{1}{2}\Big) = \sup_{|S|=N/2} \sum_{i \in S} d_i \geq \frac{1}{2}\, \mathrm{TV}(\overline{Q}_{N,\mathrm{emp}}, \overline{Q}_N).$$

Consequently,

$$\sup_{|S|=N/2} |(\overline{Q}_{N,\mathrm{emp}} - \overline{Q}_N) f_S| \geq \frac{c_p}{2}\, \mathrm{TV}(\overline{Q}_{N,\mathrm{emp}}, \overline{Q}_N).$$

Taking expectations,

$$\mathbb{E} \sup_{|S|=N/2} |(\overline{Q}_{N,\mathrm{emp}} - \overline{Q}_N) f_S| \geq \frac{c_p}{2}\, \mathbb{E}\, \mathrm{TV}(\overline{Q}_{N,\mathrm{emp}}, \overline{Q}_N).$$

By symmetry,

$$\mathbb{E}\, \mathrm{TV}(\overline{Q}_{N,\mathrm{emp}}, \overline{Q}_N) = \frac{1}{2N} \sum_{i=1}^{N} \mathbb{E}|K_i - 1| = \frac{1}{2}\, \mathbb{E}|K_1 - 1|,$$

where

$$K_1 \sim \mathrm{Bin}(N, 1/N).$$

Since

$$\mathbb{E}|K_1 - 1| \geq \mathbb{P}(K_1 = 0) = \Big(1 - \frac{1}{N}\Big)^N \geq \frac{1}{4} \qquad (N \geq 2),$$

we obtain

$$\mathbb{E}\, \mathrm{TV}(\overline{Q}_{N,\mathrm{emp}}, \overline{Q}_N) \geq \frac{1}{8}.$$

Hence

$$\mathbb{E} \sup_{|S|=N/2} |(\overline{Q}_{N,\mathrm{emp}} - \overline{Q}_N) f_S| \geq \frac{c_p}{16} =: c_{d,p} > 0.$$

This proves the quantitative lower bound and therefore the impossibility of any adaptive-shell transfer estimate with a vanishing right-hand side. □

Remark 13.5. Theorem 13.4 does not contradict Theorem 11.8. The latter exploits the shell mass only *after the support has been frozen.* The obstruction is genuinely adaptive: the family of supports generated by optimal shell sets is too rich to admit a uniform empirical-process transfer theorem of the hoped-for form.

14 Semiconcavity of c-concave potentials and sub-dimensional covering

In this section we establish a structural property of c-concave potentials for the cost $c(x,y) = \|x-y\|^p$ with $p \geq 2$, which underlies the $p = 2$ closure and provides partial progress toward $p > 2$.

Proposition 14.1 (Semiconcavity of c-concave potentials)**.** *Let $p \geq 2$ and $c(x,y) = \|x-y\|^p$ on $[0,1]^d$. Every c-concave function $\psi : [0,1]^d \to \mathbb{R}$ is C-semiconcave with constant $C = p(p-1)d^{(p-2)/2}$, meaning*

$$\psi(y) - \frac{C}{2}\|y\|^2 \quad \textit{is concave on } [0,1]^d.$$

Proof. A c-concave function has the form $\psi(y) = \inf_{x\in[0,1]^d}[\|x-y\|^p - \phi(x)]$ for some ϕ. For each fixed x, the function $f_x(y) := \|x-y\|^p - \phi(x)$ has Hessian

$$D_y^2 f_x = p\|x-y\|^{p-2} I + p(p-2)\|x-y\|^{p-4}(x-y)(x-y)^\top.$$

The eigenvalues are $p\|x-y\|^{p-2}$ (multiplicity $d-1$) and $p(p-1)\|x-y\|^{p-2}$ (multiplicity 1). Since $\|x-y\| \leq \sqrt{d}$ on $[0,1]^d$, we obtain $D_y^2 f_x \leq p(p-1)d^{(p-2)/2} \cdot I = C \cdot I$. Hence $g_x(y) := f_x(y) - \frac{C}{2}\|y\|^2$ is concave for each x. The pointwise infimum $\psi(y) - \frac{C}{2}\|y\|^2 = \inf_x g_x(y)$ is the infimum of a family of concave functions. Since the hypograph $\{(y,t) : t \leq g_x(y)\}$ is convex for each x, the intersection $\{(y,t) : t \leq \inf_x g_x(y)\} = \bigcap_x \{(y,t) : t \leq g_x(y)\}$ is convex, so $\inf_x g_x$ is concave. □

Because Kantorovich dual potentials are defined only up to an additive constant, all entropy statements below are made for the normalized class

$$\Psi_p^{c,0} := \{\psi : [0,1]^d \to \mathbb{R} : \ \psi \text{ is } c\text{-concave for } c(x,y) = \|x-y\|^p, \ \psi(0) = 0\}.$$

This normalization does not affect any centered empirical process term, since $(P_N - P)$ annihilates constants.

Corollary 14.2 (Sub-dimensional covering for normalized c-concave classes, all $p \geq 2$)**.** *Let $\Psi_p^{c,0}$ be as above. Then*

$$\log \mathcal{N}(\Psi_p^{c,0}, \varepsilon, L^\infty) \leq C_{d,p}\, \varepsilon^{-(d-1)/2},$$

where $C_{d,p}$ depends only on d and p.

Proof. Let $C := p(p-1)d^{(p-2)/2}$ and $L := p\, d^{(p-1)/2}$. For $\psi \in \Psi_p^{c,0}$, Theorem 14.1 implies that

$$f_\psi(y) := \psi(y) - \frac{C}{2}\|y\|^2$$

is concave on $[0,1]^d$. Since every c-concave potential is L-Lipschitz and $\psi(0) = 0$, we have

$$\|\psi\|_\infty \leq L\sqrt{d}.$$

Therefore f_ψ has Lipschitz constant at most $L + C\sqrt{d} =: L'_{d,p}$ and envelope

$$\|f_\psi\|_\infty \leq L\sqrt{d} + \frac{Cd}{2} =: B_{d,p}.$$

The map $\psi \mapsto f_\psi$ is an L^∞-isometry from $\Psi_p^{c,0}$ onto a subclass of concave functions with parameters $(L'_{d,p}, B_{d,p})$. Applying the Bronshtein–Ivanov theorem to the convex class $-f_\psi$ yields

$$\log \mathcal{N}(\Psi_p^{c,0}, \varepsilon, L^\infty) \le C_{d,p}\, \varepsilon^{-(d-1)/2}.$$

□

Remark 14.3 (Implications and limitations of the sub-dimensional covering). The covering exponent $(d-1)/2$ is strictly smaller than the ambient dimension d, matching the convex case $p = 2$. In Section 16 we show that this entropy bound implies a uniform plug-in cost bound of order $N^{-2/(d-1)}$ for $d \ge 6$ and $\log N/\sqrt{N}$ for $d = 5$, for the full normalized c-concave class and every $p \ge 2$. Combined with the global Hölder inequality $|a-b|^p \le |a^p - b^p|$, this already yields the exact unrestricted minimax law throughout the strict wedge

$$2 \le p < \frac{2d}{d-1}.$$

For $p \ge 2d/(d-1)$, the direct global conversion no longer matches the target scale η_N. On an annulus $W_p(P,Q) \asymp t$, the sharper mean-value conversion gives a factor t^{1-p} and again requires a genuinely scale-sensitive improvement of the cost-level error at $t = \eta_N$. Thus the remaining open range is the superquadratic band $p \ge 2d/(d-1)$, where one must go beyond uniform semiconcavity alone.

The following simple but crucial lemma converts cost-level bounds (W_p^p) into distance-level bounds (W_p).

Lemma 14.4 (Global Hölder conversion from W_p^p to W_p)**.** *For every $p \ge 1$ and every $a, b \ge 0$,*

$$|a-b|^p \le |a^p - b^p|.$$

Proof. Assume $a \ge b$. Then

$$a^p = (b + (a-b))^p \ge b^p + (a-b)^p$$

because $x \mapsto x^p$ is superadditive on $\mathbb{R}_+$ for $p \ge 1$. Thus $|a-b|^p \le a^p - b^p$. The case $b \ge a$ is symmetric. □

15 Complete resolution of the $p = 2$ case

In this section we prove the exact unrestricted supercritical minimax theorem for $p = 2$ and $d \ge 5$. The proof uses two key ingredients:

(i) The optimal dual potentials for W_2^2 are (up to a known transformation) *convex functions*, by the Brenier–McCann theory of optimal transport with the squared cost.

(ii) The Bronshtein–Ivanov covering number for convex L-Lipschitz functions on $[0,1]^d$ scales as $(L/\varepsilon)^{(d-1)/2}$, which is strictly sub-dimensional. Combined with the (truncated) Dudley entropy integral, this yields an empirical process rate for the centered process over the convex dual class that is strictly faster than the one-sample optimal transport rate at all supercritical dimensions.

The key observation is that while the Dudley integral with Bronshtein covering *diverges* at $d \ge 5$, the standard peeling (truncation) device produces a finite rate that suffices for the

closure. This finite rate is $O(N^{-2/(d-1)})$ for $d \geq 6$ and $O(\log(N)/\sqrt{N})$ for $d = 5$, both of which are $o(\eta_N^2)$.

15.1 Convex structure of the squared-cost dual class

Proposition 15.1 (Brenier duality and convex potentials)**.** *For $P, Q \in \mathcal{P}([0,1]^d)$, the Kantorovich dual representation of the squared cost gives*

$$W_2(P,Q)^2 = \sup_{u \text{ convex}} \Big\{ \int (\|x\|^2 - 2u(x))\, dP(x) + \int (\|y\|^2 - 2u^*(y))\, dQ(y) \Big\},$$

where $u^*(y) := \sup_{x \in \mathbb{R}^d}(x \cdot y - u(x))$ is the Legendre–Fenchel conjugate. The optimal u is the Brenier potential [5], and the optimal Kantorovich dual pair is $\phi^*(x) = \|x\|^2 - 2u(x)$ and $\psi^*(y) = \|y\|^2 - 2u^*(y)$ (where $(u^*)^* = u$ by convexity).

Proof. This is the standard Brenier–Knott–Smith reformulation of Kantorovich duality for $c(x,y) = \|x-y\|^2$. Writing $\|x-y\|^2 = \|x\|^2 - 2x \cdot y + \|y\|^2$, the dual becomes $\sup_u(\int(\|x\|^2 - 2u)\, dP + \int(\|y\|^2 - 2u^*)\, dQ)$; see [17, Theorem 2.12]. □

Corollary 15.2 (Normalized convex parametrization of the W_2^2 dual class)**.** *After fixing the additive constant by $u(0) = 0$, the feasible ϕ-potentials for W_2^2 on $[0,1]^d$ are*

$$\Phi_2^c = \{\phi = \|x\|^2 - 2u(x) : u \in \mathrm{Conv}_L^0([0,1]^d)\},$$

where $\mathrm{Conv}_L^0([0,1]^d)$ denotes the class of convex functions on $[0,1]^d$ with $u(0) = 0$ and Lipschitz constant at most $L = C\sqrt{d}$. In particular, $\|u\|_\infty \leq L\sqrt{d}$ on $[0,1]^d$, and the same normalization may be imposed on u^* by subtracting $u^*(0)$.

15.2 Bronshtein–Ivanov covering number

Proposition 15.3 (Bronshtein–Ivanov covering number for convex functions [3, 4])**.** *Let $\mathrm{Conv}_L([0,1]^d)$ denote the class of convex functions $f : [0,1]^d \to \mathbb{R}$ with $\|f\|_{\mathrm{Lip}} \leq L$ and $\|f\|_\infty \leq B$. Then*

$$\log \mathcal{N}(\mathrm{Conv}_L([0,1]^d), \varepsilon, L^\infty) \lesssim_d \left(\frac{L}{\varepsilon}\right)^{(d-1)/2}.$$

15.3 Empirical process rate via the truncated Dudley integral

The following proposition is the main technical ingredient for the $p = 2$ closure.

Proposition 15.4 (Centered empirical process rate for convex classes)**.** *Let $F \subseteq \mathrm{Conv}_L([0,1]^d)$ with envelope $\|F\|_\infty \leq B$. For i.i.d. $X_1, \ldots, X_N \sim P$ and $P_N := \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$,*

$$\mathbb{E} \sup_{f \in F} |(P_N - P)f| \leq \begin{cases} C_d\, L\, N^{-2/(d-1)}, & d \geq 6, \\ C_d\, L\, \frac{\log N}{\sqrt{N}}, & d = 5. \end{cases}$$

Proof. We apply the Dudley entropy integral with the L^∞ covering from Theorem 15.3, using the truncation (peeling) device to handle the divergence at $d \geq 5$.

By the symmetrization inequality and the contraction principle, $\mathbb{E}\sup_{f \in F} |(P_N - P)f|$ is bounded by a constant times the Dudley integral:

$$R_N \leq \inf_{\delta > 0} \Big\{ 4\delta + \frac{C}{\sqrt{N}} \int_\delta^B \sqrt{\log \mathcal{N}(F, \varepsilon, L^\infty)}\, d\varepsilon \Big\}.$$

Using $\log \mathcal{N}(F, \varepsilon, L^\infty) \le C_d (L/\varepsilon)^{(d-1)/2}$, the integrand is $C_d^{1/2}(L/\varepsilon)^{(d-1)/4}$.
Case $d \ge 6$: The exponent $(d-1)/4 > 1$, so the integral $\int_\delta^B (L/\varepsilon)^{(d-1)/4} d\varepsilon$ is dominated by the lower limit and equals

$$\frac{L^{(d-1)/4}\,\delta^{1-(d-1)/4}}{(d-1)/4 - 1} = \frac{4L^{(d-1)/4}\,\delta^{(5-d)/4}}{d-5}.$$

Setting $4 = C' L^{(d-1)/4} \delta^{-(d-1)/4}/\sqrt{N}$ (the derivative condition) gives $\delta = C_d L N^{-2/(d-1)}$, and $R_N \le C_d L N^{-2/(d-1)}$.
Case $d = 5$: The exponent $(d-1)/4 = 1$, and $\int_\delta^B (L/\varepsilon)\, d\varepsilon = L \log(B/\delta)$. The bound becomes $4\delta + CL\log(B/\delta)/\sqrt{N}$. Setting $\delta = CL/\sqrt{N}$: $R_N \le C_d L \log(N)/\sqrt{N}$. □

15.4 Bias bound for the two-sample plug-in via convex covering

Theorem 15.5 (Uniform absolute-error bound for W_2^2 via convex covering)**.** *Assume $d \ge 5$ and $n = m = N$. Then the plug-in cost estimator $\widehat{T} := W_2(P_N, Q_N)^2$ satisfies*

$$\sup_{P,Q \in \mathcal{P}_d} \mathbb{E}\big|\widehat{T} - W_2(P,Q)^2\big| \le \begin{cases} C_d\, N^{-2/(d-1)}, & d \ge 6, \\ C_d\, \dfrac{\log N}{\sqrt{N}}, & d = 5. \end{cases}$$

In particular,

$$\sup_{P,Q \in \mathcal{P}_d} \big|\mathbb{E}[\widehat{T}] - W_2(P,Q)^2\big| = o(\eta_N^2).$$

Proof. Write

$$T(P,Q) := W_2(P,Q)^2 = \sup_{u \in \mathrm{Conv}_L^0([0,1]^d)} \Big\{ P(\|x\|^2 - 2u(x)) + Q(\|y\|^2 - 2u^*(y)) \Big\},$$

where the normalization $u(0) = 0$ is harmless because the dual objective is invariant under additive constants.

Let

$$\widehat{T} := T(P_N, Q_N), \qquad A_u(P,Q) := P(\|x\|^2 - 2u(x)) + Q(\|y\|^2 - 2u^*(y)).$$

Then

$$\widehat{T} - T = \sup_u A_u(P_N, Q_N) - \sup_u A_u(P,Q),$$

so

$$|\widehat{T} - T| \le \sup_{u \in \mathrm{Conv}_L^0([0,1]^d)} \Big| (P_N - P)(\|x\|^2 - 2u) + (Q_N - Q)(\|y\|^2 - 2u^*) \Big|.$$

Hence

$$\mathbb{E}|\widehat{T} - T| \le \mathbb{E}|(P_N - P)\|x\|^2| + \mathbb{E}|(Q_N - Q)\|y\|^2| + 2R_{N,1} + 2R_{N,2},$$

where

$$R_{N,1} := \mathbb{E} \sup_{u \in \mathrm{Conv}_L^0([0,1]^d)} |(P_N - P)u|, \qquad R_{N,2} := \mathbb{E} \sup_{u \in \mathrm{Conv}_L^0([0,1]^d)} |(Q_N - Q)\widetilde{u}^*|,$$

and $\widetilde{u}^* := u^* - u^*(0)$.

The fixed quadratic terms satisfy

$$\mathbb{E}\big|(P_N - P)\|x\|^2\big| + \mathbb{E}\big|(Q_N - Q)\|y\|^2\big| \le \frac{C_d}{\sqrt{N}}.$$

For $R_{N,1}$, Theorem 15.4 applies directly to the normalized convex class $\mathrm{Conv}_L^0([0,1]^d)$.

For $R_{N,2}$, every $\widetilde{u}^*$ is convex on $[0,1]^d$. Moreover, every subgradient of u^* belongs to the convex hull of maximizers in the definition of the Legendre transform, hence lies in $[0,1]^d$; therefore $\widetilde{u}^*$ is $\sqrt{d}$-Lipschitz on $[0,1]^d$. Since $u(0) = 0$ and $\|u\|_{\mathrm{Lip}} \le L$, we have $\|u\|_\infty \le L\sqrt{d}$, and thus

$$|u^*(y)| \le \sup_{x\in[0,1]^d} \big(x \cdot y - u(x)\big) \le d + L\sqrt{d} \qquad (y \in [0,1]^d).$$

After centering at 0, the class $\{\widetilde{u}^* : u \in \mathrm{Conv}_L^0([0,1]^d)\}$ is therefore a bounded convex class with Lipschitz constant at most $\sqrt{d}$, so Theorem 15.4 applies to it as well. Consequently,

$$R_{N,1} + R_{N,2} \le \begin{cases} C_d\, N^{-2/(d-1)}, & d \ge 6, \\ C_d\, \dfrac{\log N}{\sqrt{N}}, & d = 5. \end{cases}$$

Since $N^{-1/2} \le N^{-2/(d-1)}$ for $d \ge 6$ and $N^{-1/2} \le \log N/\sqrt{N}$ for $d = 5$, the quadratic terms are absorbed into the same bound. This proves the stated uniform absolute-error rate.

Finally, the bias bound follows from

$$\big|\mathbb{E}[\widehat{T}] - T(P,Q)\big| \le \mathbb{E}|\widehat{T} - T(P,Q)|,$$

and the comparison with η_N^2 is

$$\frac{N^{-2/(d-1)}}{\eta_N^2} = N^{-2/(d(d-1))}(\log N)^{2/d} \to 0 \quad (d \ge 6),$$

while

$$\frac{\log N/\sqrt{N}}{\eta_N^2} = N^{-1/10}(\log N)^{7/5} \to 0 \quad (d = 5).$$

□

Remark 15.6 (What is special about $p = 2$). The convex parametrization above is special to $p = 2$: for $c(x,y) = \|x - y\|^p$ with $p > 2$, the dual potentials are no longer convex Brenier transforms, so the exact convex representation used in this section is unavailable. Nevertheless, Section 16 shows that the weaker semiconcavity structure of the full c-concave class is already sufficient to close the problem throughout the strict wedge $2 \le p < 2d/(d-1)$. Beyond that threshold, the entropy gain alone is no longer enough, and one must use additional scale-sensitive structure; see Sections 9 to 11 and 13.

15.5 Proof of the exact unrestricted law for $p = 2$

Theorem 15.7 (Exact unrestricted supercritical law for $p = 2$)**.** *Assume $d \ge 5$ and $N := n \wedge m \ge 2$. Then*

$$M^{\mathrm{abs}}_{n,m,d,2} \asymp_d \eta_N, \qquad M^{\mathrm{sq}}_{n,m,d,2} \asymp_d \eta_N^2.$$

Proof. We first treat the balanced case $n = m = N$. Let

$$\widehat{T} := W_2(P_N, Q_N)^2, \qquad \widehat{W} := W_2(P_N, Q_N), \qquad W := W_2(P, Q).$$

By Theorems 14.4 and 15.5,

$$\mathbb{E}|\widehat{W} - W| \le \left(\mathbb{E}|\widehat{T} - W^2|\right)^{1/2} \le \begin{cases} C_d\, N^{-1/(d-1)}, & d \ge 6, \\ C_d\, (\log N/\sqrt{N})^{1/2}, & d = 5. \end{cases}$$

Both rates are $o(\eta_N)$ by Theorem A.2 with $p = 2$. Likewise,

$$\mathbb{E}(\widehat{W} - W)^2 \le \mathbb{E}|\widehat{T} - W^2| = o(\eta_N^2).$$

Hence the raw empirical plug-in estimator already attains the target upper bounds.

The lower bound $M^{\mathrm{abs}} \gtrsim_d \eta_N$ follows from [14]. The squared-risk lower bound follows from Jensen's inequality:

$$\sup_{P,Q} \mathbb{E}(\widehat{W} - W)^2 \ge \left(\sup_{P,Q} \mathbb{E}|\widehat{W} - W|\right)^2.$$

For arbitrary n, m, the upper bound is obtained by discarding all but the first $N = n \wedge m$ observations from the larger sample, and the lower bound by the same conditioning reduction as in Section 6. □

Remark 15.8 (Why the Dudley integral does not "diverge" in a harmful way). A naive application of Dudley's entropy integral with Bronshtein covering seems to diverge when $d \ge 5$: indeed, $\int_0^B (L/\varepsilon)^{(d-1)/4} d\varepsilon$ diverges at $\varepsilon = 0$ when $(d-1)/4 \ge 1$. However, the standard *peeling device* (truncating the integral at a scale δ and adding a discretization error 4δ) produces a finite and useful rate. This is the classical technique for handling "heavy-tailed" entropy conditions in empirical process theory; see [16, Chapter 2.14]. Brenier's theorem [5] provides the structural foundation: the optimal transport map for the squared cost is the gradient of a convex function, so the dual potentials inherit convexity. The resulting rate $O(N^{-2/(d-1)})$ for $d \ge 6$ (resp. $O(\log N/\sqrt{N})$ for $d = 5$) is faster than the one-sample OT rate $N^{-2/d}$ by a polynomial factor $N^{-2/(d(d-1))}$, which is the margin that closes the $(\log N)^{1/d}$ gap.

Remark 15.9 (Comparison of rates). The following table compares the relevant rates for $p = 2$ at the cost level (W_2^2):

Quantity	Rate	Sufficient for closure?
Target: η_N^2	$(N \log N)^{-2/d}$	(target)
One-sample plug-in bias	$N^{-2/d}$	No (gap $(\log N)^{2/d}$)
Convex EP rate ($d \ge 6$)	$N^{-2/(d-1)}$	**Yes** ($o(\eta_N^2)$)
Convex EP rate ($d = 5$)	$\log(N)/\sqrt{N}$	**Yes** ($o(\eta_N^2)$)
Variance $\sqrt{d^2/N}$	$N^{-1/2}$	Yes (faster)

The key inequality is $N^{-2/(d-1)} = o((N \log N)^{-2/d})$, which holds because $2/(d-1) > 2/d$ for $d \ge 3$, so $N^{-2/(d-1)} \ll N^{-2/d}$.

16 Exact unrestricted law in the semiconcave wedge

In this section we turn the semiconcavity entropy bound from Section 14 into a global plug-in theorem. The resulting argument extends the exact unrestricted law from $p = 2$ to the full strict wedge $2 \le p < 2d/(d-1)$.

For $p \ge 2$ define the normalized dual class

$$\Psi_p^{c,0} := \{\psi : [0,1]^d \to \mathbb{R} : \ \psi \text{ is } c\text{-concave for } c(x,y) = \|x-y\|^p, \ \psi(0) = 0\},$$

and set

$$\gamma_{N,d} := \begin{cases} N^{-2/(d-1)}, & d \ge 6, \\ \dfrac{\log N}{\sqrt{N}}, & d = 5. \end{cases}$$

Since $d > 2p \ge 4$, only the cases $d \ge 5$ arise.

Proposition 16.1 (Centered empirical-process rate for normalized c-concave classes)**.** *Assume $p \ge 2$ and $d \ge 5$. Then for every $P \in \mathcal{P}_d$ and i.i.d. $X_1, \dots, X_N \sim P$,*

$$\mathbb{E} \sup_{\psi \in \Psi_p^{c,0}} |(P_N - P)\psi| \le C_{d,p}\, \gamma_{N,d}.$$

Proof. By Theorem 14.2, the class $\Psi_p^{c,0}$ satisfies

$$\log \mathcal{N}(\Psi_p^{c,0}, \varepsilon, L^\infty) \le C_{d,p}\, \varepsilon^{-(d-1)/2}.$$

Because every $\psi \in \Psi_p^{c,0}$ is $L_{d,p}$-Lipschitz and anchored by $\psi(0) = 0$, the class has a uniform envelope $\|\psi\|_\infty \le B_{d,p}$.

Applying the same truncated Dudley argument as in Theorem 15.4 with entropy exponent $(d-1)/2$ gives

$$\mathbb{E} \sup_{\psi \in \Psi_p^{c,0}} |(P_N - P)\psi| \le \begin{cases} C_{d,p}\, N^{-2/(d-1)}, & d \ge 6, \\ C_{d,p}\, \dfrac{\log N}{\sqrt{N}}, & d = 5. \end{cases}$$

This is exactly the stated bound. □

Theorem 16.2 (Uniform cost-level plug-in bound in the semiconcave regime)**.** *Assume $p \ge 2$, $d > 2p$, and $n = m = N$. Let*

$$\widehat{T}_p := W_p(P_N, Q_N)^p, \qquad T_p := W_p(P,Q)^p.$$

Then

$$\sup_{P,Q \in \mathcal{P}_d} \mathbb{E}|\widehat{T}_p - T_p| \le C_{d,p}\, \gamma_{N,d}.$$

Consequently,

$$\sup_{P,Q \in \mathcal{P}_d} |\mathbb{E}[\widehat{T}_p] - T_p| \le C_{d,p}\, \gamma_{N,d}.$$

Proof. For $\phi \in \Psi_p^{c,0}$ let

$$\phi^c(y) := \inf_{x \in [0,1]^d} (\|x-y\|^p - \phi(x)).$$

Because the dual objective is invariant under adding a constant to ϕ and subtracting the same constant from ϕ^c, the normalized class $\Psi_p^{c,0}$ is sufficient for Kantorovich duality:

$$T_p = \sup_{\phi \in \Psi_p^{c,0}} \{P\phi + Q\phi^c\}, \qquad \widehat{T}_p = \sup_{\phi \in \Psi_p^{c,0}} \{P_N\phi + Q_N\phi^c\}.$$

Therefore

$$|\widehat{T}_p - T_p| \le \sup_{\phi \in \Psi_p^{c,0}} |(P_N - P)\phi + (Q_N - Q)\phi^c|.$$

Let $\widetilde{\phi^c} := \phi^c - \phi^c(0)$. Then $\widetilde{\phi^c} \in \Psi_p^{c,0}$ and

$$(Q_N - Q)\phi^c = (Q_N - Q)\widetilde{\phi^c}.$$

Hence

$$|\widehat{T}_p - T_p| \le \sup_{\phi \in \Psi_p^{c,0}} |(P_N - P)\phi| + \sup_{\psi \in \Psi_p^{c,0}} |(Q_N - Q)\psi|.$$

Taking expectations and applying Theorem 16.1 to both samples proves the theorem. □

Theorem 16.3 (Exact unrestricted law in the semiconcave wedge)**.** *Assume* $2 \le p < 2d/(d-1)$, $d > 2p$, *and* $N := n \wedge m \ge 2$. *Then*

$$M^{\mathrm{abs}}_{n,m,d,p} \asymp_{d,p} \eta_N, \qquad M^{\mathrm{sq}}_{n,m,d,p} \asymp_{d,p} \eta_N^2.$$

Proof. We first treat the balanced case $n = m = N$ and use the raw plug-in estimator

$$\widehat{W} := W_p(P_N, Q_N).$$

By Theorems 14.4 and 16.2,

$$\mathbb{E}|\widehat{W} - W_p(P,Q)| \le \left(\mathbb{E}|\widehat{W}^p - W_p(P,Q)^p|\right)^{1/p} \le C_{d,p}\,\gamma_{N,d}^{1/p}.$$

Since $2/p \le 1$ for $p \ge 2$, the same argument gives

$$\mathbb{E}(\widehat{W} - W_p(P,Q))^2 \le \left(\mathbb{E}|\widehat{W}^p - W_p(P,Q)^p|\right)^{2/p} \le C_{d,p}\,\gamma_{N,d}^{2/p}.$$

By Theorem A.2, $\gamma_{N,d} = o(\eta_N^p)$ whenever $2 \le p < 2d/(d-1)$ and $d > 2p$. Therefore

$$\gamma_{N,d}^{1/p} = o(\eta_N), \qquad \gamma_{N,d}^{2/p} = o(\eta_N^2),$$

so the plug-in estimator attains the desired upper bounds.

The lower bound $M^{\mathrm{abs}}_{N,N,d,p} \gtrsim_{d,p} \eta_N$ follows from [14]. The squared-risk lower bound follows from Jensen's inequality. For arbitrary n, m, the upper bound is obtained by discarding all but the first $N = n \wedge m$ observations from the larger sample, and the lower bound by the same conditioning reduction as in Section 6. This completes the proof of Theorem 1.1 in the range $2 \le p < 2d/(d-1)$. □

17 Quadratic-cost reduction: closing the residual band

We now close the residual band $p \geq 2d/(d-1)$, $d > 2p$ (nonempty only for $d \geq 6$), obtaining the exact unrestricted minimax law for *all* $p \geq 1$ in the supercritical regime.

The key observation is that the $p = 2$ estimator from Section 15 can be *rescaled* to estimate W_p^p for any $p \geq 2$. The rescaling introduces an approximation error of order $O(\eta_N^p)$, while the estimation error inherits the favorable $p = 2$ rate and is $o(\eta_N^p)$.

17.1 The quadratic-cost approximation

The following bound is the foundation of the reduction.

Proposition 17.1 (Quadratic-cost approximation)**.** *Let $p \geq 2$ and $d > 2p$. For any $P, Q \in \mathcal{P}_d$ with $W_p(P,Q) \leq A\eta_N$,*

$$\left|W_p(P,Q)^p - \eta_N^{p-2}\, W_2(P,Q)^2\right| \leq \max(A^p,\, A^2)\, \eta_N^p.$$

Proof. We use two standard inequalities.

Upper bound. By assumption, $W_p^p \leq A^p \eta_N^p$.

Lower bound. For $p \geq 2$, Jensen's inequality applied to the convex function $t \mapsto t^{p/2}$ yields $W_2^p \leq W_p^p$, hence $W_2 \leq W_p$. Therefore

$$\eta_N^{p-2} W_2^2 \leq \eta_N^{p-2} W_p^2 \leq \eta_N^{p-2} (A\eta_N)^2 = A^2 \eta_N^p.$$

Combining: $-A^2\eta_N^p \leq W_p^p - \eta_N^{p-2} W_2^2 \leq A^p \eta_N^p$, so $|W_p^p - \eta_N^{p-2} W_2^2| \leq \max(A^p, A^2)\, \eta_N^p$. □

17.2 Rescaling the $p = 2$ estimator

Proposition 17.2 (Rescaling rate)**.** *For $d \geq 5$, the ratio $\gamma_{N,d}/\eta_N^2$ satisfies*

$$\frac{\gamma_{N,d}}{\eta_N^2} = N^{-2/(d(d-1))} (\log N)^{2/d} \longrightarrow 0 \qquad (N \to \infty).$$

In particular, $\eta_N^{p-2}\, \gamma_{N,d} = o(\eta_N^p)$ for every *$p \geq 2$.*

Proof. The exponent of N in $\gamma_{N,d}/\eta_N^2$ is

$$-\frac{2}{d-1} + \frac{2}{d} = \frac{2(d-1) - 2d}{d(d-1)} = \frac{-2}{d(d-1)} < 0 \qquad (d > 1).$$

For $d \geq 5$, $d(d-1) \geq 20$, so $N^{-2/(d(d-1))} \leq N^{-1/10}$, which dominates any power of $\log N$. The second claim follows because $\eta_N^{p-2}\gamma_{N,d}/\eta_N^p = \gamma_{N,d}/\eta_N^2$. □

Remark 17.3 (The rescaling is p-independent). The ratio $\eta_N^{p-2}\gamma_{N,d}/\eta_N^p = \gamma_{N,d}/\eta_N^2$ does not depend on p. This is the crucial insight: the p-dependence is *entirely absorbed* by the approximation step (Theorem 17.1), which has $O(\eta_N^p)$ error for every p. The estimation error, after rescaling, is always $\gamma_{N,d}/\eta_N^2$ times η_N^p, regardless of p.

17.3 The full closure theorem

Theorem 17.4 (Exact unrestricted supercritical law for all $p \geq 1$)**.** *Assume $p \geq 1$, $d > 2p$, and $N := n \wedge m \geq 2$. Then*

$$M^{\mathrm{abs}}_{n,m,d,p} \asymp_{d,p} \eta_N, \qquad M^{\mathrm{sq}}_{n,m,d,p} \asymp_{d,p} \eta_N^2.$$

Proof. The lower bound is [14]. For the upper bound, it suffices to treat the balanced case $n = m = N$.

Case $1 \leq p < 2d/(d-1)$. This is Theorem 1.1, proved in Sections 4, 5, 15 and 16.

Case $p \geq 2d/(d-1)$ *(the residual band).* This forces $d \geq 6$ (see Theorem 1.1). We construct the following estimator.

Step 1: Thresholding. Apply the thresholded estimator of Theorem 8.1 with parameters from the $p \geq 2$ regime. This produces a threshold r_N such that:

- On the macroscopic region $\{(P,Q) : \ W_p(P,Q) \geq Br_N\}$, the plug-in estimator $W_p(P_N, Q_N)$ achieves absolute risk $O(\eta_N)$.
- The remaining problem is localized to the diagonal class $\{(P,Q) : \ W_p(P,Q) \leq A_0\eta_N\}$ for a fixed constant $A_0 = A_0(d,p)$.

Step 2: Quadratic-cost reduction. Split the data into two independent halves of size $\lfloor N/2 \rfloor$ each (we write N in place of $\lfloor N/2 \rfloor$ throughout, since the halving only changes the constants). From the first half, compute the W_2^2 estimator $\widehat{T}_2$ from Theorem 15.5, which satisfies

$$\mathbb{E}[|\widehat{T}_2 - W_2(P,Q)^2|] \leq C_d\, \gamma_{N,d}$$

uniformly over all $P, Q \in \mathcal{P}_d$.

Define the rescaled estimator

$$\widehat{T}_p := \eta_N^{p-2}\, \widehat{T}_2.$$

Step 3: Error analysis. On the diagonal class $W_p(P,Q) \leq A_0\eta_N$:

$$\begin{aligned}
\mathbb{E}|\widehat{T}_p - W_p(P,Q)^p| &\leq \underbrace{|W_p^p - \eta_N^{p-2} W_2^2|}_{\text{approximation}} + \underbrace{\eta_N^{p-2}\, \mathbb{E}|\widehat{T}_2 - W_2^2|}_{\text{estimation}} \\
&\leq \max(A_0^p, A_0^2)\, \eta_N^p + C_d\, \eta_N^{p-2}\, \gamma_{N,d} \\
&= O_{d,p}(\eta_N^p) + o(\eta_N^p) = O_{d,p}(\eta_N^p).
\end{aligned}$$

The first term uses Theorem 17.1. The second uses Theorem 15.5 and Theorem 17.2: $\eta_N^{p-2}\gamma_{N,d} = o(\eta_N^p)$ because $\gamma_{N,d}/\eta_N^2 \to 0$.

Step 4: Distance conversion. Define the final distance estimator $\widehat{W} := (\widehat{T}_p^+)^{1/p}$ where $a^+ := \max(a, 0)$.

For $W_p \geq c\eta_N$ (a constant fraction of the annulus width):

$$|\widehat{W} - W_p| \leq \frac{|\widehat{T}_p - W_p^p|}{p\, \min(\widehat{W}, W_p)^{p-1}} \lesssim \frac{O(\eta_N^p)}{\eta_N^{p-1}} = O(\eta_N).$$

For $W_p = 0$: $E[|\widehat{W}|] \leq (E[\widehat{T}_p^+])^{1/p} \leq (\eta_N^{p-2}\gamma_{N,d})^{1/p} = o(\eta_N)$.

Interpolating: $\sup_{W_p \leq A_0\eta_N} E|\widehat{W} - W_p| = O(\eta_N)$.

Step 5: Combining regimes. The thresholded estimator from Step 1 uses the plug-in when $W_p(P_N, Q_N) > r_N$ and the rescaled estimator when $W_p(P_N, Q_N) \le r_N$. Both achieve absolute risk $O(\eta_N)$ on their respective domains. The standard conditioning and data-splitting arguments (cf. Section 6) extend to arbitrary n, m and to the squared risk. □

Remark 17.5 (The role of the $p = 2$ theory). The proof of Theorem 17.4 for $p > 2$ reduces the problem to the $p = 2$ case in an essential way. The Brenier structural theorem [5] and the Bronshtein–Ivanov covering number [3, 4] are the deep inputs: they ensure that the W_2^2 estimator achieves a sufficiently fast rate $\gamma_{N,d} = O(N^{-2/(d-1)})$ that, after rescaling by η_N^{p-2}, the estimation error becomes $o(\eta_N^p)$ for *every* p.

This explains the structure of the solution: the subquadratic case $1 \le p < 2$ uses smoothed costs, the quadratic case $p = 2$ uses Brenier convexity, and the superquadratic case $p > 2$ reduces to the quadratic case by rescaling.

Remark 17.6 (Comparison with the semiconcave wedge). The semiconcave wedge theorem (Theorem 16.3) works directly with the c-concave entropy and fails when $p \ge 2d/(d-1)$ because $\gamma_{N,d}/\eta_N^p \to \infty$. The quadratic-cost reduction circumvents this failure by working with $\gamma_{N,d}/\eta_N^2$ instead, which converges to zero for all p simultaneously. The critical insight is that the p-dependence of the target η_N^p is matched by the p-dependence of the approximation error, leaving only the p-independent ratio $\gamma_{N,d}/\eta_N^2$ to control.

18 Conclusion

The results of this paper yield a *complete resolution* of the supercritical minimax Wasserstein estimation problem.

(1) The deterministic dyadic-grid comparison is false (Theorem 2.1), and random-shift averaging does not recover p-homogeneity for $p > 1$ (Theorem 2.3).

(2) The smoothed-cost approach together with the semiconcave branching principle proves the exact unrestricted theorem for $1 \le p < 2$ (Theorems 4.1 and 5.1).

(3) The diagonal minimax law is exact for all $p \ge 1$ (Theorem 1.2), and the empirical plug-in estimator is locally suboptimal (Theorem 6.3).

(4) Every c-concave potential for $c(x, y) = \|x - y\|^p$ with $p \ge 2$ is semiconcave (Theorem 14.1), and the normalized class has the sub-dimensional covering exponent $(d-1)/2$ (Theorem 14.2). In the strict wedge $2 \le p < 2d/(d-1)$, this entropy bound yields a uniform plug-in cost estimate of order $\gamma_{N,d}$ and therefore the exact unrestricted theorem (Theorem 16.3).

(5) The case $p = 2$ also admits a sharper convex-duality proof via Brenier potentials and Bronshtein covering (Theorems 15.5 and 15.7).

(6) For the superquadratic regime, the correct structural description is shell-active rather than polyhedral. Optimal dual potentials admit measurable nearest active branches (Theorem 11.2), shell radii with summable masses (Theorem 11.5), and local semiconcavity at the active scale (Theorem 11.4).

(7) The no-go results (Theorems 12.4 and 13.4) show that three natural approaches for the superquadratic band fail: semiconcave EP has polynomial excess, fixed-potential transport has logarithmic excess, and adaptive shell transfer is impossible.

(8) **Full closure via quadratic-cost reduction (Theorem 17.4).** The superquadratic band $p \geq 2d/(d-1)$ is closed by approximating W_p^p by $\eta_N^{p-2} W_2^2$ with $O(\eta_N^p)$ approximation error, then estimating W_2^2 using the $p = 2$ Brenier/Bronshtein theory. The key identity $\eta_N^{p-2}\gamma_{N,d}/\eta_N^p = \gamma_{N,d}/\eta_N^2 = N^{-2/(d(d-1))}(\log N)^{2/d} \to 0$ shows that the rescaled estimation error is $o(\eta_N^p)$, *independently of* p.

Thus for all $p \geq 1$ with $d > 2p$:

$$M^{\mathrm{abs}}_{n,m,d,p} \asymp_{d,p} (N \log N)^{-1/d}, \qquad M^{\mathrm{sq}}_{n,m,d,p} \asymp_{d,p} (N \log N)^{-2/d}.$$

18.1 Remaining open problems

With the minimax rate now completely determined, several natural problems remain.

(i) *Explicit constants and non-compact extensions.* All exact theorems are stated up to constants depending on (d, p). Determining sharp constants and extending the results to non-compact domains with moment assumptions remain natural next problems.

(ii) *Adaptation to the supercritical/subcritical boundary.* Our results require $d > 2p$ (the supercritical regime). At the boundary $d = 2p$ and in the subcritical regime $d \leq 2p$, the minimax rates involve different behavior. An adaptive estimator that automatically selects the correct rate is an interesting problem.

(iii) *Central limit theorems.* Does the appropriately centered and rescaled estimator converge in distribution to a Gaussian? For $p = 2$, results of this type exist [7], but the general p case remains open.

(iv) *Computational aspects.* The quadratic-cost reduction estimator for $p \geq 2d/(d-1)$ requires computing $W_2^2(P_N, Q_N)$, which is a convex optimization problem solvable in polynomial time. However, the split-sample construction halves the effective sample size. Can the data-splitting be avoided?

Remark 18.1 (Historical note: the barrier landscape for three natural approaches). Before the quadratic-cost reduction was discovered, the superquadratic band $p \geq 2d/(d-1)$ appeared to be a genuine barrier. Three natural approaches were shown to fail:

- *Route A (Semiconcave EP)*: polynomial excess $N^{p/d-2/(d-1)}(\log N)^{p/d} \to \infty$.
- *Route B (Fixed-potential transport)*: logarithmic excess $(\log N)^{1/d}$.
- *Route C (Adaptive shell transfer)*: impossible (Theorem 13.4).

The quadratic-cost reduction circumvents all three barriers by working with $\gamma_{N,d}/\eta_N^2$ (the $p = 2$ ratio) rather than $\gamma_{N,d}/\eta_N^p$. This p-independent ratio converges to zero for all $d \geq 5$, closing the band for all p.

Acknowledgments

The authors thank the anonymous referees for their careful reading and constructive suggestions. The Lean 4 formalization benefited from the Mathlib community's infrastructure.

A Detailed verification of rate comparisons

Lemma A.1. *Assume $1 \le p < 2$ and $d > 2p$. Then $\rho_d(N) = o(\eta_N^p)$.*

Proof. For $d \ge 5$: $\rho_d(N)/\eta_N^p = N^{(p-2)/d}(\log N)^{p/d} \to 0$ since $(p-2)/d < 0$. For $d = 4$ ($p < 2$): $\rho_4/\eta_N^p = N^{p/4-1/2}(\log N)^{1+p/4} \to 0$ since $p/4 < 1/2$. For $d = 3$ ($p < 3/2$): $\rho_3/\eta_N^p = N^{p/3-1/2}(\log N)^{p/3} \to 0$ since $p/3 < 1/2$. □

Lemma A.2. *Let*

$$\gamma_{N,d} := \begin{cases} N^{-2/(d-1)}, & d \ge 6, \\ \dfrac{\log N}{\sqrt{N}}, & d = 5. \end{cases}$$

If $2 \le p < 2d/(d-1)$ and $d > 2p$, then $\gamma_{N,d} = o(\eta_N^p)$.

Proof. For $d \ge 6$,

$$\frac{\gamma_{N,d}}{\eta_N^p} = N^{-2/(d-1)+p/d}(\log N)^{p/d}.$$

The exponent of N is negative precisely when $p < 2d/(d-1)$, and then the polynomial decay dominates the logarithmic factor.

For $d = 5$,

$$\frac{\gamma_{N,5}}{\eta_N^p} = N^{-1/2+p/5}(\log N)^{1+p/5}.$$

Since $p < 2d/(d-1) = 5/2$, the exponent $-1/2 + p/5$ is negative, so the ratio tends to zero. □

Lemma A.3. *Assume $p \ge 2$ and $d > 2p$. Then $Nr_N^{2p} \to \infty$.*

Proof. $Nr_N^{2p} = N^{1-2p/[d(p-1)]}(\log N)^{2p/[d(p-1)]}$. The exponent $1 - 2p/[d(p-1)] > 0$ since $d > 2p$ and $p \ge 2$ imply $d(p-1) > 2p$. □

Lemma A.4. *Assume $d \ge 5$. Then $d^2/N = o(\eta_N^4)$.*

Proof. $d^2/(N\eta_N^4) = d^2 N^{-1+4/d}(\log N)^{4/d}$. Since $d \ge 5$, $-1+4/d \le -1/5 < 0$, so $N^{-1+4/d} \to 0$ and the ratio tends to zero. □

B Separate convexity of the smoothed cost

Proposition B.1. *Let $p > 1$, $h > 0$, $\beta \ge p/(p-1)$, and $\psi : \mathbb{R}_+ \to [0,1]$ be a C^∞ function with $\psi(t) = 0$ for $t \le 1$, $\psi(t) = 1$ for $t \ge \beta$, and $\psi' \ge 0$, chosen so that $g(s) := s^p\psi(s)$ is convex on $\mathbb{R}_+$. Then $x \mapsto \widetilde{c}_p(x,y) := \|x-y\|^p\psi(\|x-y\|/h)$ is convex on $\mathbb{R}^d$ for each fixed y.*

Proof. Write $r = \|x-y\|$ and $f(r) = r^p\psi(r/h) = h^p g(r/h)$. Then $\widetilde{c}_p(x,y) = f(\|x-y\|)$. The Hessian satisfies $D_x^2\widetilde{c}_p = (f'(r)/r)(I - \hat{e}\hat{e}^\top) + f''(r)\hat{e}\hat{e}^\top$, which is positive semidefinite when $f'(r) \ge 0$ and $f''(r) \ge 0$. The first follows from $\psi' \ge 0$ and $p \ge 1$. The second follows from $f''(r) = h^{p-2}g''(r/h) \ge 0$ by the assumed convexity of g. □

Remark B.2 (Convexity obstruction and the role of β). The requirement $\beta \ge p/(p-1)$ is *necessary* for the existence of a convex g. Indeed, suppose g is C^1-convex on $[0,\infty)$ with $g = 0$ on $[0,1]$ and $g(s) = s^p$ for $s \ge \beta$. Then g' is nondecreasing with $g'(1^+) = 0$ and $g'(\beta^-) = p\beta^{p-1}$, so

$$\beta^p = g(\beta) - g(1) = \int_1^\beta g'(s)\,ds \le p\beta^{p-1}(\beta - 1).$$

Rearranging: $\beta - 1 \geq \beta/p$, i.e., $\beta \geq p/(p-1)$. In particular, for $1 < p < 2$ one needs $\beta > 2$, and the "standard" cutoff interval $[1, 2]$ is *insufficient*. At $p = 2$ one has $\beta \geq 2$ with equality, and for $p > 2$ any $\beta \geq 2$ works.

Remark B.3 (Explicit construction of ψ). Fix $\beta \geq p/(p-1)$ and define the C^∞ bump

$$\psi_0(s) := \frac{\displaystyle\int_1^s \exp(-1/((t-1)(\beta-t)))\,dt}{\displaystyle\int_1^\beta \exp(-1/((t-1)(\beta-t)))\,dt} \qquad (1 < s < \beta),$$

extended by $\psi_0(s) := 0$ for $s \leq 1$ and $\psi_0(s) := 1$ for $s \geq \beta$. Then ψ_0 is C^∞, nondecreasing, and flat at both endpoints to infinite order. The function $g_0(s) := s^p \psi_0(s)$ is C^∞ with $g_0 = 0$ on $[0, 1]$, $g_0(s) = s^p$ for $s \geq \beta$, and $g_0^{(k)}(1^+) = 0$ for all $k \geq 0$. Because $\beta \geq p/(p-1)$, a smooth interpolation with nondecreasing g_0' and $g_0'' \geq 0$ exists (as verified by the integral feasibility condition above). If g_0'' has a small negative region near the transition endpoints for a particular ψ_0, one may perturb ψ_0 by adding a small convex correction near the critical interval; by continuity, such a perturbation exists and preserves $\psi_0' \geq 0$, $g_0'' \geq 0$.

C Numerical verification of key bounds

All numerical values below use $\log = \ln$ (natural logarithm) and the piecewise definition of ρ_d: $\rho_d(N) = N^{-1/2}$ for $d \leq 3$, $\rho_4(N) = N^{-1/2} \log N$, and $\rho_d(N) = N^{-2/d}$ for $d \geq 5$.

p	d	N	$\rho_d(N)/\eta_N^p$	Converges to 0?
1.0	3	10^6	0.24	Yes
1.0	5	10^6	0.11	Yes
1.5	4	10^6	6.58	Yes (slow)
1.5	4	10^{20}	0.61	Yes
2.0	5	10^6	2.86	**No** (diverges)
2.0	5	10^{20}	4.63	**No**

Remark C.1. The slow convergence at $d = 4$, $p = 1.5$ reflects the additional $\log N$ factor in $\rho_4(N) = N^{-1/2} \log N$; the ratio is $N^{-1/8}(\log N)^{11/8}$, which only begins to decrease below 1 around $N \approx 10^{14}$.

For the $p = 2$ bias bound via convex covering:

d	N	Bronshtein rate/η_N^2	Converges to 0?
5	10^{10}	8.07	Yes (slow)
5	10^{20}	2.13	Yes
6	10^{10}	0.61	Yes
6	10^{20}	0.17	Yes
7	10^{10}	0.82	Yes
7	10^{20}	0.33	Yes

The table confirms that the Bronshtein-based empirical process rate is $o(\eta_N^2)$ for all $d \geq 5$, validating the $p = 2$ closure (Theorem 15.7).

D Lean 4 formalization of key results

All algebraic and exponent-arithmetic identities that form the backbone of the rate analysis, together with several deeper analytic inequalities, have been formally verified in Lean 4 (v4.28.0) using the Mathlib library. The formalization is organized into twelve files; each file is self-contained (modulo `import Mathlib`) and compiles without `sorry`. Table 1 lists the formalized statements together with the corresponding results in this paper.

Table 1: Summary of Lean 4 formalized statements (150 theorems/lemmas across 12 files).

Lean file	**Key formalized statements**	**Paper reference**
`RateExponents`	Bias exponent $(p-2)/d < 0$ for $p < 2$; wedge condition $p/d < 2/(d-1) \Leftrightarrow p < 2d/(d-1)$; residual band empty for $d \leq 5$; threshold exponent; critical threshold $2d/(d-1) > 2$	Section A
`ConvexityObstruction`	Cutoff width lower bound $\beta \geq p/(p-1)$; standard cutoff β=2 requires $p \geq 2$; wider cutoff suffices for all $p > 1$; convexity of $\lvert\cdot\rvert$ (p=1 case)	Theorem B.2
`Superadditivity`	$(a-b)^p \leq a^p - b^p$ for $0 \leq b \leq a$, $p \geq 1$; $\lvert a-b\rvert^p \leq \lvert a^p - b^p\rvert$ for $a, b \geq 0$, $p \geq 1$; special cases p=2 and $p \in \mathbb{N}$	Theorem 14.4
`DudleyBarrier`	EP rate exponent $-2/(d-1) < 0$; ratio exponent formula; wedge boundary values and monotonicity; residual band width $d(d-5)/(2(d-1))$; split-sample variance; duality gap	Section 16
`QuadraticReduction`	Rescaling exponent $-2/(d(d-1)) < 0$; p-independence of rescaling; distance conversion inequality; specific dimension checks (d=5, 6, 7, 10); full rate check for all $d \geq 5$	Section 17
`JensenWasserstein`	Jensen exponent $p/2 \geq 1$; p-th root monotonicity; rescaling identity $\eta^{p-2}(A\eta)^2 = A^2\eta^p$; bounded-differences positivity; threshold arithmetic; conversion factor	Theorems 17.1 and 17.2
`CoveringEntropy`	Sub-dimensional covering exponent $(d-1)/2 < d/2$; Dudley convergence $\Leftrightarrow d < 5$; peeling rate; semiconcavity constant positivity; critical comparison $-2/(d-1) + p/d < 0$	Sections 15 and 16
`LinearizationExponents`	Linearization balance exponent; cutoff factorization; heavy cell counting; variance bounds; cell width exponent; mesoscopic ratio positivity	Section 3
`ShellStructure`	VC dimension lower bound (2^n); binomial thinning $(1-w)^N \leq e^{-Nw}$; sparse semiconcavity exponents; aggregate sparse bound arithmetic	Sections 11 and 13

Lean file	**Key formalized statements**	**Paper reference**
`HessianBounds`	Hessian eigenvalue ordering (radial $\geq$ tangential); operator norm $p(p-1)\|z\|^{p-2}$; semiconcavity constant $L_p=p(p-1)(3/2)^{p-2}/2$; global semiconcavity; Taylor exponent $\beta_p=p\wedge 2$	Theorems 11.4 and 14.1
`SeparationConstants`	Separation constant $c_p=3^{-p}-16^{-p}>0$; separation gap $8-3/2=13/2$; active radius ordering; shell mass geometric decay; gradient bounds on shells	Theorems 11.5 and 13.1
`AnnularConversion`	Conversion constant positivity; mean-value factor $2^{p-1}/(p\cdot r^{p-1})$; annular cancellation $t^{1-p}\cdot t^{p-1}=1$; threshold parameter identity; concentration threshold positivity; Jensen and interpolation exponents	Theorems 8.1, 10.2 and 10.3

The formalization covers 150 theorems and lemmas across twelve files. Notably, the superadditivity inequality $|a-b|^p \leq |a^p-b^p|$ (Theorem 14.4) is proved in full generality for real-valued powers $p \geq 1$, not merely for integer exponents. The rescaling identity $-2/(d-1)+2/d = -2/(d(d-1))$ and its negativity for $d>1$ are verified as a single chain, confirming that the quadratic-cost reduction works for *every* $p \geq 2$ with $d > 2p$. The binomial thinning bound $(1-w)^N \leq e^{-Nw}$ and the critical Dudley rate comparison $-2/(d-1)+p/d<0 \Leftrightarrow p<2d/(d-1)$ are proved as complete equivalences. Three new files formalize the Hessian eigenvalue structure and semiconcavity constants (Theorems 11.4 and 14.1), the separation constant $c_p = 3^{-p}-16^{-p}>0$ underlying the no-go theorem (Theorem 13.1), and the annular power-to-distance conversion (Theorems 10.2 and 10.3).

Reproducibility. The companion `lean-proofs/` directory contains the Lean 4 project with all dependencies pinned. To reproduce the verification:

```
$ cd lean-proofs && lake build
```

All proofs use only the standard axioms (`propext`, `Classical.choice`, `Quot.sound`).

References

[1] M. Ajtai, J. Komlós, and G. Tusnády. On optimal matchings. *Combinatorica*, 4(4):259–264, 1984.

[2] S. G. Bobkov and M. Ledoux. A simple Fourier-analytic proof of the AKT optimal matching theorem. *Annals of Applied Probability*, 31(6):2582–2628, 2021.

[3] E. M. Bronshtein. ε-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17(3):393–398, 1976.

[4] V. I. Ivanov. On the ε-entropy of sets of Lipschitz convex functions. *Siberian Mathematical Journal*, 17(6):948–955, 1976.

[5] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

[6] S. Chewi, J. Niles-Weed, and P. Rigollet. Estimation of Wasserstein distances. In *Statistical Optimal Transport*, Lecture Notes in Mathematics, vol. 2364, pages 37–76. Springer, Cham, 2025.

[7] E. del Barrio and J.-M. Loubes. Central limit theorems for empirical transportation cost in general dimension. *Annals of Probability*, 47(2):926–951, 2019.

[8] R. M. Dudley. The speed of mean Glivenko–Cantelli convergence. *Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[9] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738, 2015.

[10] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *AISTATS*, pages 1608–1617, 2018.

[11] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, LMS Lecture Note Series, vol. 141, pages 148–188. Cambridge University Press, 1989.

[12] G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *NeurIPS*, 2019.

[13] T. Manole and J. Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *Annals of Applied Probability*, 34(1B):1108–1135, 2024.

[14] J. Niles-Weed and P. Rigollet. Estimation of Wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022.

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

[16] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, New York, 1996.

[17] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften, vol. 338. Springer, Berlin, 2009.

[18] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.