

Exact Fixed-Dimensional Minimax Theory for Wasserstein Distance Estimation

Churilov M. V.

churilovm1305@gmail.com

Abstract

We establish the unrestricted fixed-dimensional minimax theory for estimating the two-sample Wasserstein distance $W_p(P, Q)$ over the full class of Borel probability measures on $[0, 1]^d$. For every $p \geq 1$, every $d > 2p$, and arbitrary sample sizes n, m with $N := n \wedge m$, we prove that the exact minimax absolute and squared risks are

$$(N \log N)^{-1/d} \quad \text{and} \quad (N \log N)^{-2/d}.$$

The argument is entirely Euclidean. First, we show that the lower-envelope scale is already exact on a full diagonal neighborhood and that empirical optimal transport remains locally suboptimal there. Second, we identify the critical finite model by reducing the balanced supercritical problem to the Euclidean problem on the dyadic grid with $|X_J| \asymp N \log N$ atoms, and we determine the generic off-diagonal geometry through full positive-distance neighborhoods, mixed-scale two-ball classes, and translated thin-annulus examples. Third, a critical linearization of the p -cost separates a parametric first-order affine layer from normalized second-order residuals; semiconcave branching bounds and support-complexity eliminations then reduce the unresolved part of the problem to a finite nearest-neighbor macrograph. Finally, that terminal layer is closed by three different mechanisms: uniform semiconcavity for $p \geq 2$, canonical corner kernels plus finite switch-envelope closure for $1 < p < 2$, and cone reduction with exact dyadic prefactor restoration for $p = 1$. A one-sided lower bound with one exact reference marginal then propagates the balanced theorem to arbitrary sample sizes.

Along the way we obtain exact local laws at the diagonal, sharp envelopes on separated Euclidean product neighborhoods, a mixed-scale phase diagram on full two-ball classes, translated off-diagonal families with logarithmic plug-in gaps, and critical-grid elimination theorems for support-sparse and dyadically thin geometries.

Keywords. Wasserstein distance; minimax estimation; optimal transport; high-dimensional statistics; dyadic partitions; mixed-scale neighborhoods.

MSC 2020. Primary 62G05; Secondary 49Q22, 60F25.

1 Introduction

Let $d \in \mathbb{N}$, $p \geq 1$, and let \mathcal{P}_d denote the set of Borel probability measures on $[0, 1]^d$. Given $P, Q \in \mathcal{P}_d$, the p -Wasserstein distance is

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

From independent samples

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

the goal is to estimate the scalar functional $W_p(P, Q)$.

The empirical approximation problem and the direct functional-estimation problem are different. By the triangle inequality,

$$|W_p(P_n, Q_m) - W_p(P, Q)| \leq W_p(P_n, P) + W_p(Q_m, Q),$$

so the empirical plug-in estimator inherits the classical curse of dimensionality from one-sample Wasserstein approximation. In the supercritical regime $d > 2p$, this yields the benchmark rate $N^{-1/d}$, where $N := n \wedge m$, by the sharp one-sample theory of Weed and Bach [9]. What is not settled by those facts alone is whether the scalar functional $W_p(P, Q)$ can be estimated strictly faster than empirical plug-in over the unrestricted class \mathcal{P}_d .

Before the results proved here, the strongest general lower bound in fixed dimension was the balanced supercritical bound of Niles–Weed and Rigollet,

$$\inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} |\widehat{W} - W_p(P, Q)| \gtrsim (N \log N)^{-1/d} \quad (d > 2p, n = m = N),$$

which left a logarithmic gap to the plug-in upper bound $N^{-1/d}$ [7]. The problem continued to be recorded as open in later expository accounts; see, for example, [1]. The sharp off-diagonal plug-in theory of Manole and Niles-Weed [6] also indicates that the hard part of the unrestricted problem is not in smooth positive-baseline classes, but in singular or weakly separated configurations.

This paper closes that gap completely. For every $p \geq 1$, every $d > 2p$, and arbitrary sample sizes n, m with $N = n \wedge m$, we prove that

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

Equivalently, in fixed dimension the supercritical two-sample Wasserstein problem exhibits a genuine sample-size enlargement from N to $N \log N$, and this enlargement is exact for the scalar functional $W_p(P, Q)$ itself.

Informal main theorem. Assume $d > 2p$ and set $N := n \wedge m$. In the balanced case $n = m = N$, there exists a constant $A_0(d, p) > 0$ such that on the full Euclidean local class

$$\mathcal{L}_{A_0, N} = \{(P, Q) \in \mathcal{P}_d^2 : W_p(P, Q) \leq A_0(N \log N)^{-1/d}\},$$

the exact minimax rates are

$$\inf_{\widehat{W}} \sup_{(P, Q) \in \mathcal{L}_{A_0, N}} \mathbb{E} |\widehat{W} - W_p(P, Q)| \asymp (N \log N)^{-1/d},$$

$$\inf_{\widehat{W}} \sup_{(P, Q) \in \mathcal{L}_{A_0, N}} \mathbb{E} (\widehat{W} - W_p(P, Q))^2 \asymp (N \log N)^{-2/d},$$

and the empirical plug-in estimator still has worst-case squared risk

$$\sup_{(P, Q) \in \mathcal{L}_{A_0, N}} \mathbb{E} (W_p(P_N, Q_N) - W_p(P, Q))^2 \gtrsim N^{-2/d}$$

on the same local class. The local theorem globalizes: for arbitrary n, m with $N := n \wedge m$,

$$\inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} |\widehat{W} - W_p(P, Q)| \asymp_{d,p} (N \log N)^{-1/d},$$

and

$$\inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} (\widehat{W} - W_p(P, Q))^2 \asymp_{d,p} (N \log N)^{-2/d}.$$

Proof strategy and structural outputs. The proof has four main layers.

- (i) *Exact diagonal geometry.* We show that the lower-envelope scale already lives on the actual Euclidean model. The diagonal theorem is proved on the full neighborhood $\mathcal{L}_{A,N}$, not on a surrogate tree or on a specially engineered subclass.
- (ii) *Critical discrete model and off-diagonal phase diagram.* At the sharp scale, the balanced supercritical problem is equivalent to the Euclidean problem on the dyadic grid with $|X_J| \asymp N \log N$ atoms. Away from the diagonal, full positive-distance product neighborhoods and full two-ball classes reveal the generic smooth and mixed-scale regimes, while translated thin-annulus classes show that off-diagonal logarithmic plug-in gaps persist only under exact first-order cancellation.
- (iii) *Reduction of the singular layer.* A critical linearization of the p -cost isolates a first-order affine layer and normalized second-order residuals. The affine layer is already parametric after pooling, a semiconcave branching principle removes all dual-regular residual families, and support-complexity elimination theorems show that support-sparse or dyadically thin macro-geometries are already subcritical on the critical grid. What remains is a finite nearest-neighbor macrograph.
- (iv) *Closure of the terminal models.* The terminal nearest-neighbor layer is closed in three distinct ways: by uniform semiconcavity when $p \geq 2$, by canonical corner kernels together with finite switch-envelope closure when $1 < p < 2$, and by a cone reduction with exact dyadic prefactor restoration when $p = 1$. A one-sided lower bound with one exact reference marginal then propagates the balanced theorem to arbitrary sample sizes.

Beyond the main minimax theorem, the argument yields exact local laws at the diagonal, sharp envelopes on separated Euclidean product neighborhoods, a mixed-scale phase diagram for full two-ball classes, translated off-diagonal subclasses with logarithmic plug-in gaps, and elimination theorems for semiconcave branching, support-sparse families, and dyadically thin critical-grid geometries. Together these results identify which Euclidean configurations are genuinely hard and which are already below the target scale.

Organization of the paper. Section 2 records the known minimax picture and the main structural results. Section 3 gives an exact paired Euclidean model on which direct estimation already beats plug-in, while Section 4 explains why exact one-scale reductions are rigid and too restrictive to capture the unrestricted supercritical difficulty. Section 5 proves the exact local theorem and the local plug-in gap, and Section 6 gives the critical-grid reduction. Sections 7 to 9 analyze the positive-distance regimes: separated classes, full two-ball neighborhoods, and translated thin annuli. Section 10 proves the known-reference lower bound used for sample-size imbalance. Section 11 develops the critical linearization, semiconcavity, support-complexity,

and finite-offset reduction package. Section 12 closes the balanced supercritical regime for $p > 1$, and Section 13 closes the boundary case $p = 1$. Finally, Section 14 propagates the balanced theorem to arbitrary sample sizes and records the full fixed-dimensional minimax picture.

2 Main theorem and structural results

For $n, m \geq 1$, define the minimax absolute and squared risks

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

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

Throughout, $N := n \wedge m$.

2.1 Previously known minimax envelopes

Theorem 2.1 (Previously known unrestricted envelopes). *Fix $d \geq 1$, $p \geq 1$, and let $N := n \wedge m$.*

(i) *If $d < 2p$, then*

$$\mathfrak{M}_{n,m,d,p}^{\text{abs}} \asymp N^{-1/2}, \quad \mathfrak{M}_{n,m,d,p}^{\text{sq}} \asymp N^{-1}.$$

(ii) *If $d = 2p$, then*

$$N^{-1/2} \lesssim \mathfrak{M}_{n,m,d,p}^{\text{abs}} \lesssim N^{-1/2} (\log N)^{1/p},$$

$$N^{-1} \lesssim \mathfrak{M}_{n,m,d,p}^{\text{sq}} \lesssim N^{-1} (\log N)^{2/p}.$$

(iii) *If $d > 2p$, then*

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

(iv) *If $d > 2p$ and $n = m = N$, then*

$$(N \log N)^{-1/d} \lesssim \mathfrak{M}_{N,N,d,p}^{\text{abs}} \lesssim N^{-1/d},$$

$$(N \log N)^{-2/d} \lesssim \mathfrak{M}_{N,N,d,p}^{\text{sq}} \lesssim N^{-2/d}.$$

Proof. The upper bounds follow from plug-in together with the sharp one-sample empirical Wasserstein bounds of Weed and Bach [9] and the compact-support estimates of Fournier and Guillin [2]. The generic lower bound $N^{-1/2}$ is obtained by restricting one marginal to a fixed atom and reducing the problem to estimating a Bernoulli mean. The balanced supercritical lower bound $(N \log N)^{-1/d}$ is due to Niles–Weed and Rigollet [7, Theorem 11 and the discussion after it]. \square

2.2 Main results

Theorem 2.2 (Exact unrestricted supercritical minimax law). *Assume $p \geq 1$, $d > 2p$, and $N := n \wedge m \geq 2$. Then*

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

Theorem 2.3 (Exact fixed-dimensional minimax picture). *Fix $d \geq 1$, $p \geq 1$, and $N := n \wedge m \geq 2$.*

(i) *If $d < 2p$, then*

$$\mathfrak{M}_{n,m,d,p}^{\text{abs}} \asymp N^{-1/2}, \quad \mathfrak{M}_{n,m,d,p}^{\text{sq}} \asymp N^{-1}.$$

(ii) *If $d = 2p$, then*

$$N^{-1/2} \lesssim \mathfrak{M}_{n,m,d,p}^{\text{abs}} \lesssim N^{-1/2}(\log N)^{1/p},$$

$$N^{-1} \lesssim \mathfrak{M}_{n,m,d,p}^{\text{sq}} \lesssim N^{-1}(\log N)^{2/p}.$$

(iii) *If $d > 2p$, then*

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

The proof of Theorem 2.2 is built from the structural Euclidean results stated below. For $d > 2p$ and $n = m = N$, write

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

For $A > 0$, define the local Euclidean class

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

Theorem 2.4 (Exact local law at the diagonal). *Assume $d > 2p$ and $n = m = N$. There exists $A_0 = A_0(d, p) > 0$ such that*

$$\inf_{\widehat{W}} \sup_{(P,Q) \in \mathcal{L}_{A_0,N}} \mathbb{E}|\widehat{W} - W_p(P, Q)| \asymp_{d,p} \eta_N,$$

$$\inf_{\widehat{W}} \sup_{(P,Q) \in \mathcal{L}_{A_0,N}} \mathbb{E}(\widehat{W} - W_p(P, Q))^2 \asymp_{d,p} \eta_N^2.$$

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

$$\sup_{(P,Q) \in \mathcal{L}_{A,N}} \mathbb{E}(W_p(P_N, Q_N) - W_p(P, Q))^2 \gtrsim_{d,p} N^{-2/d}.$$

The first point of Theorem 2.4 is that the lower-envelope scale $(N \log N)^{-1/d}$ is already exact on a full Euclidean neighborhood of the diagonal. The second point is that plug-in remains locally suboptimal by a full logarithmic factor.

The next result identifies the correct finite-state model behind the balanced supercritical problem. Let $J = J(N, L)$ be defined by

$$2^{Jd} \leq LN \log N < 2^{(J+1)d},$$

and let X_J be the set of centers of the level- J dyadic cubes. Denote by $\mathfrak{G}_{N,L,d,p}^{\text{abs}}$ and $\mathfrak{G}_{N,L,d,p}^{\text{sq}}$ the minimax absolute and squared risks when both marginals are restricted to $\mathcal{P}(X_J)$.

Theorem 2.5 (Critical-grid reduction). *Assume $d > 2p$ and $n = m = N$. For every fixed $L \geq 1$,*

$$\mathfrak{G}_{N,L,d,p}^{\text{abs}} \leq \mathfrak{M}_{N,N,d,p}^{\text{abs}} \leq \mathfrak{G}_{N,L,d,p}^{\text{abs}} + C_d 2^{-J},$$

$$\mathfrak{G}_{N,L,d,p}^{\text{sq}} \leq \mathfrak{M}_{N,N,d,p}^{\text{sq}} \leq 2\mathfrak{G}_{N,L,d,p}^{\text{sq}} + C_d 2^{-2J}.$$

In particular, for every sufficiently large fixed L , the unrestricted balanced supercritical problem is equivalent, up to constants, to the same Euclidean estimation problem on a dyadic grid with $|X_J| \asymp N \log N$ atoms.

Away from the diagonal, the picture becomes geometric. Fix a small absolute constant $\gamma := 1/20$, set

$$z_\star := \left(\frac{1}{4}, \frac{1}{2}, \dots, \frac{1}{2}\right), \quad X_\star := \{u \in \mathbb{R}^d : \|u\|_2 \leq \gamma\}, \quad Y_\star := e_1 + X_\star,$$

and for $0 < r \leq 1/4$ define

$$X_r := z_\star + rX_\star, \quad Y_r := z_\star + rY_\star, \quad \mathcal{C}_r := \mathcal{P}(X_r) \times \mathcal{P}(Y_r).$$

Theorem 2.6 (Full positive-distance neighborhoods). *Assume $d \geq 5$ and $p > 1$. There exist constants $0 < c_{d,p} \leq C_{d,p} < \infty$ such that for every $0 < r \leq 1/4$,*

$$c_{d,p} r (N \log N)^{-2/d} \leq \inf_{\widehat{W}} \sup_{(P,Q) \in \mathcal{C}_r} \mathbb{E} |\widehat{W} - W_p(P,Q)| \leq C_{d,p} r N^{-2/d}.$$

Thus, on separated product neighborhoods, the smooth-cost exponent $N^{-2/d}$ replaces the local diagonal scale $(N \log N)^{-1/d}$. The next regime is genuinely mixed: the two marginals are separated by a positive baseline r , but each is itself confined to a much thinner ball of radius $s \ll r$.

Let

$$B := \{u \in \mathbb{R}^d : \|u\|_2 \leq 1\}, \quad z_\diamond := \left(\frac{1}{4}, \frac{1}{2}, \dots, \frac{1}{2}\right),$$

and define

$$\mathcal{C}_{r,s} := \mathcal{P}(z_\diamond + sB) \times \mathcal{P}(z_\diamond + re_1 + sB), \quad 0 < s \leq r/20.$$

Theorem 2.7 (Mixed-scale phase diagram on full two-ball classes). *Assume $d \geq 5$, $p > 1$, and $n = m = N$. There exist $\kappa_\star = \kappa_\star(d, p) \in (0, 1/20]$ and constants $0 < c_{d,p} \leq C_{d,p} < \infty$ such that for every $0 < r \leq 1/4$ and $0 < s \leq \kappa_\star r$,*

$$\begin{aligned} c_{d,p} \max\left\{sN^{-1/2}, \frac{s^2}{r} \eta_N^2\right\} &\leq \inf_{\widehat{W}} \sup_{(P,Q) \in \mathcal{C}_{r,s}} \mathbb{E} |\widehat{W} - W_p(P,Q)| \\ &\leq C_{d,p} \left(sN^{-1/2} + \frac{s^2}{r} N^{-2/d}\right). \end{aligned}$$

Equivalently, the critical width is

$$s_{\text{crit}}(N, r) \asymp rN^{2/d-1/2}.$$

Below that width, the full class is generically parametric with risk $\asymp sN^{-1/2}$; above it, the full class is generically nonparametric with risk between $(s^2/r)(N \log N)^{-2/d}$ and $(s^2/r)N^{-2/d}$.

Theorem 2.8 (A nongeneric off-diagonal logarithmic gap). *Fix $0 < r \leq 1/10$, assume $d \geq 2$, $p \geq 1$, and let $n = m = N$. For all sufficiently large N , there exists an exact Euclidean class $\mathcal{A}_{N,r}^{\text{off}} \subset \mathcal{P}_d^2$ such that every pair $(P, Q) \in \mathcal{A}_{N,r}^{\text{off}}$ obeys*

$$r \leq W_p(P, Q) \leq r + C_{d,p} r N^{-2/(d-1)},$$

while a direct estimator achieves

$$\sup_{(P,Q) \in \mathcal{A}_{N,r}^{\text{off}}} \mathbb{E}(\widehat{W} - W_p(P, Q))^2 \lesssim_{d,p} r^2 N^{-4/(d-1)} (\log N)^{-1/p},$$

and the empirical plug-in estimator satisfies

$$\sup_{(P,Q) \in \mathcal{A}_{N,r}^{\text{off}}} \mathbb{E}(W_p(P_N, Q_N) - W_p(P, Q))^2 \gtrsim_{d,p} r^2 N^{-4/(d-1)}.$$

Thus direct estimation can still beat empirical optimal transport far from the diagonal, but only on a smaller translated geometry in which the first-order term cancels exactly.

For arbitrary sample sizes, a lower-bound obstruction already survives in a one-sided experiment with one exact reference marginal; see Theorem 2.9 and Section 10.

Proposition 2.9 (One exact reference already preserves the lower envelope). *Assume $d > 2p$. There exist constants $A_1 = A_1(d, p)$, $c_1 = c_1(d, p) > 0$, and $n_1 = n_1(d, p)$ such that for all $n \geq n_1$ there is a reference measure $Q_n \in \mathcal{P}_d$ with*

$$\inf_{\widehat{W}} \sup_{P \in \mathcal{L}_{A_1, n}^{\rightarrow}(Q_n)} \mathbb{E}|\widehat{W}(X_1, \dots, X_n; Q_n) - W_p(P, Q_n)| \gtrsim_{d,p} (n \log n)^{-1/d},$$

with the analogous lower bound $\gtrsim_{d,p} (n \log n)^{-2/d}$ for squared loss. Consequently, any balanced supercritical upper bound propagates to arbitrary sample sizes by discarding surplus observations.

Remark 2.10. Theorems 2.4 to 2.9, together with the reduction and closure theorems of Sections 11 to 14, yield Theorem 2.2. The diagonal theorem supplies the sharp local scale; the grid theorem identifies the correct finite-state model; the positive-distance and mixed-scale theorems control the generic off-diagonal geometry; and the remaining singular layer is finite and is closed separately in the ranges $p \geq 2$, $1 < p < 2$, and $p = 1$.

2.3 Proof architecture

The structural results above separate the unrestricted problem into three geometrically distinct regimes.

First, Theorem 2.4 settles the exact diagonal law on the continuum model itself. This is where the lower-envelope scale $(N \log N)^{-1/d}$ first becomes an upper bound, and it already shows that empirical optimal transport remains logarithmically suboptimal on a full Euclidean neighborhood of the diagonal.

Second, Theorems 2.5 to 2.8 identify the correct finite model and the generic off-diagonal geometry. The critical-grid theorem shows that the balanced supercritical problem is equivalent, up to constants, to the same Euclidean estimation problem on a dyadic grid with $\asymp N \log N$ atoms. The positive-distance and mixed-scale theorems then determine the smooth and mixed-scale off-diagonal regimes, while the thin-annulus construction isolates the nongeneric situation in which a logarithmic off-diagonal gap survives exact first-order cancellation.

Third, the later sections reduce and close the singular layer left by these geometric results. Critical linearization converts the unresolved part of the problem into a finite-offset expansion with a first-order affine term and normalized second-order residuals. The affine layer is already parametric, the semiconcave branching principle removes all dual-regular residual classes, and Theorem 11.4 eliminates support-sparse and dyadically thin critical-grid families. What

remains is a finite nearest-neighbor macrograph, closed by uniform semiconcavity for $p \geq 2$, by canonical corner kernels plus finite switch-envelope closure for $1 < p < 2$, and by a cone reduction with exact prefactor restoration for $p = 1$.

3 An exact paired Euclidean model

Before turning to the full local theorem, it is useful to record an explicit Euclidean class on which direct functional estimation already beats empirical plug-in. The point of this model is conceptual: it shows that the logarithmic phenomenon is genuinely Euclidean and does not require any tree surrogate.

Lemma 3.1 (Paired packing). *There exists a constant $c_d > 0$ such that the following holds. For every $M \geq 1$ and every $0 < \delta \leq c_d M^{-1/d}$, one can choose points*

$$\{x_1^+, x_1^-, \dots, x_M^+, x_M^-\} \subset [0, 1]^d$$

with

$$\|x_j^+ - x_j^-\|_2 = \delta \quad (j = 1, \dots, M),$$

and such that distinct pairs are separated by at least 5δ :

$$\min\{\|u - v\|_2 : u \in \{x_i^+, x_i^-\}, v \in \{x_j^+, x_j^-\}\} \geq 5\delta \quad (i \neq j).$$

Proof. Choose a dyadic grid with spacing 8δ in every coordinate. For c_d small enough, the grid contains at least M points in $[0, 1]^d$. Given a selected anchor z_j , set

$$x_j^- := z_j + 2\delta e_1, \quad x_j^+ := z_j + 3\delta e_1.$$

Then $\|x_j^+ - x_j^-\|_2 = \delta$, and points belonging to different pairs remain separated by at least 5δ . \square

Proposition 3.2 (Exact paired Euclidean reduction). *Fix $M \geq 1$, $p \geq 1$, and a paired configuration as in Theorem 3.1. For $\theta = (\theta_1, \dots, \theta_M) \in [-1, 1]^M$, define*

$$P_\theta := \frac{1}{2M} \sum_{j=1}^M [(1 + \theta_j)\delta_{x_j^+} + (1 - \theta_j)\delta_{x_j^-}].$$

Then for every $\theta, \eta \in [-1, 1]^M$,

$$W_p(P_\theta, P_\eta)^p = \frac{\delta^p}{2M} \sum_{j=1}^M |\theta_j - \eta_j|.$$

Equivalently, if

$$q_\theta(j, \pm) = \frac{1 \pm \theta_j}{2M}, \quad q_\eta(j, \pm) = \frac{1 \pm \eta_j}{2M},$$

then

$$W_p(P_\theta, P_\eta)^p = \delta^p \text{TV}(q_\theta, q_\eta).$$

Proof. Each measure places total mass $1/M$ on every pair $\{x_j^+, x_j^-\}$. Hence any coupling induces a circulation between the macro-pairs. If an optimal coupling transported positive mass around a cycle of distinct pairs, rerouting that mass inside each participating pair would

strictly lower the cost because inter-pair moves cost at least $(5\delta)^p$, whereas the corresponding intra-pair corrections cost only δ^p . Thus some optimal coupling is entirely localized within the pairs. Inside a fixed pair j , exactly $|\theta_j - \eta_j|/(2M)$ units of mass must cross the edge of length δ . Summing these contributions proves the formula. \square

Let

$$\mathcal{G}_{M,\delta}^{\text{pair}} := \{P_\theta : \theta \in [-1, 1]^M\} \subset \mathcal{P}_d.$$

Theorem 3.3 (Direct estimator on the paired class). *Assume $n = m = N$, let $\log N \leq C_0 \log(2M)$, where C_0 is the constant from Jiao–Han–Weissman [3], and let $\mathcal{G}_{M,\delta}^{\text{pair}}$ be the paired class above. Then there exists an estimator \widehat{W} such that*

$$\sup_{P,Q \in \mathcal{G}_{M,\delta}^{\text{pair}}} \mathbb{E} |\widehat{W} - W_p(P, Q)| \lesssim_p \delta \left(\frac{M}{N \log N} \right)^{1/(2p)},$$

$$\sup_{P,Q \in \mathcal{G}_{M,\delta}^{\text{pair}}} \mathbb{E} (\widehat{W} - W_p(P, Q))^2 \lesssim_p \delta^2 \left(\frac{M}{N \log N} \right)^{1/p}.$$

Proof. By Theorem 3.2, estimating $W_p(P_\theta, P_\eta)^p$ is exactly the same as estimating $\delta^p \text{TV}(q_\theta, q_\eta)$ on an alphabet of size $2M$. Apply the L_1 -distance estimator of Jiao, Han, and Weissman [3] to obtain an estimator \widehat{L} of $\|q_\theta - q_\eta\|_1$ with squared error $\lesssim M/(N \log N)$. Set

$$\widehat{V} := \frac{\delta^p}{2} \widehat{L}, \quad \widehat{W} := \widehat{V}^{1/p}.$$

Then

$$\sup_{\theta, \eta} \mathbb{E} (\widehat{V} - W_p(P_\theta, P_\eta)^p)^2 \lesssim_p \delta^{2p} \frac{M}{N \log N}.$$

Since $x \mapsto x^{1/p}$ is $1/p$ -Hölder on $[0, \infty)$,

$$|\widehat{W} - W_p(P_\theta, P_\eta)| \leq |\widehat{V} - W_p(P_\theta, P_\eta)^p|^{1/p},$$

which yields the stated absolute and squared risk bounds. \square

Corollary 3.4 (A strict logarithmic improvement on an exact Euclidean subclass). *Assume $d > 2p$ and $n = m = N$. Then there exists an exact Euclidean subclass $\mathcal{G}_N^{\text{pair}} \subset \mathcal{P}_d$ such that*

$$\sup_{P,Q \in \mathcal{G}_N^{\text{pair}}} \mathbb{E} |\widehat{W} - W_p(P, Q)| \lesssim_{d,p} N^{-1/d} (\log N)^{-1/(2p)}$$

for a suitable direct estimator, whereas plug-in satisfies

$$\sup_{P,Q \in \mathcal{G}_N^{\text{pair}}} \mathbb{E} |W_p(P_N, Q_N) - W_p(P, Q)| \gtrsim_{d,p} N^{-1/d}.$$

Proof. Take $M \asymp N$ and $\delta \asymp N^{-1/d}$ in Theorem 3.3. This gives the direct-estimation upper bound. For the lower bound on plug-in, apply the same argument as in Theorem 5.14 at the uniform paired point $Q_0 \in \mathcal{G}_N^{\text{pair}}$, for which the minimum nonzero support separation is $\asymp N^{-1/d}$ and the true distance is zero. \square

4 Rigidity and limitations of exact one-scale reductions

The paired construction is exact because it preserves mass pairwise. If one instead looks for a universal identity of the form $W_p^p = c \text{TV}$ on a single fixed support, the geometry becomes rigid and the resulting statistical problem is much easier than the supercritical unrestricted problem.

For a finite codebook $X = \{x_1, \dots, x_M\} \subset [0, 1]^d$, let Δ_M denote the simplex of probability vectors on M points, and write

$$F_X(q) := \sum_{i=1}^M q_i \delta_{x_i}, \quad q \in \Delta_M.$$

Proposition 4.1 (Rigidity of exact one-scale TV reductions). *Assume that there exists $c > 0$ such that*

$$W_p(F_X(q), F_X(r))^p = c \text{TV}(q, r) \quad \text{for all } q, r \in \Delta_M.$$

Then X is equilateral:

$$\|x_i - x_j\|_2^p = c \quad (i \neq j).$$

In particular, $M \leq d + 1$.

Proof. Take $q = \delta_i$ and $r = \delta_j$. Then

$$W_p(F_X(q), F_X(r))^p = \|x_i - x_j\|_2^p, \quad \text{TV}(q, r) = 1,$$

so every off-diagonal distance equals c . A classical fact of Euclidean geometry says that an equilateral set in \mathbb{R}^d has size at most $d + 1$. \square

Proposition 4.2 (Exact equilateral reduction). *Assume that $X = \{x_1, \dots, x_M\} \subset [0, 1]^d$ is equilateral with common separation Δ :*

$$\|x_i - x_j\|_2 = \Delta \quad (i \neq j).$$

Then for every $q, r \in \Delta_M$,

$$W_p(F_X(q), F_X(r))^p = \Delta^p \text{TV}(q, r).$$

Proof. The overlapping mass $\sum_i \min\{q_i, r_i\} \delta_{x_i}$ can be matched at zero cost. Every remaining unit of mass must move between distinct support points, and every such move costs exactly Δ^p . Thus the optimal cost equals Δ^p times the unmatched mass, namely $\text{TV}(q, r)$. \square

Theorem 4.3 (Fixed equilateral codebooks have the rate $N^{-1/p}$). *Let $X \subset [0, 1]^d$ be equilateral of size $M \geq 2$ and common separation Δ , and let $N = n \wedge m$. Then*

$$c_p \Delta^2 N^{-1/p} \leq \inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}(X)} \mathbb{E}(\widehat{W} - W_p(P, Q))^2 \leq C_{M,p} \Delta^2 N^{-1/p},$$

$$c'_p \Delta N^{-1/(2p)} \leq \inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}(X)} \mathbb{E}|\widehat{W} - W_p(P, Q)| \leq C'_{M,p} \Delta N^{-1/(2p)}.$$

Proof. Write $P = F_X(q)$ and $Q = F_X(r)$. By Theorem 4.2,

$$W_p(P, Q) = \Delta \text{TV}(q, r)^{1/p}.$$

For the upper bound, use the empirical histograms \widehat{q} and \widehat{r} , set $\widehat{T} := \text{TV}(\widehat{q}, \widehat{r})$, and define $\widehat{W} := \Delta \widehat{T}^{1/p}$. By Theorem 5.4,

$$\sup_{q, r \in \Delta_M} \mathbb{E}(\widehat{T} - \text{TV}(q, r))^2 \lesssim_M N^{-1}.$$

The Hölder continuity of $x \mapsto x^{1/p}$ therefore yields the displayed upper bounds.

For the lower bound, restrict to

$$r_0 = (1/2, 1/2, 0, \dots, 0), \quad q_\varepsilon = (1/2 + \varepsilon, 1/2 - \varepsilon, 0, \dots, 0).$$

Then $W_p(F_X(q_\varepsilon), F_X(r_0)) = \Delta \varepsilon^{1/p}$. Estimating the target over this two-point family reduces to distinguishing two Bernoulli laws whose Kullback–Leibler divergence is $\lesssim N\varepsilon^2$. Choosing $\varepsilon \asymp N^{-1/2}$ and applying the standard two-point lower bound gives the claims. \square

Corollary 4.4 (Exact deterministic one-scale reductions are statistically subcritical). *Assume $d > 2p$, and suppose a fixed support family $X \subset [0, 1]^d$ satisfies*

$$W_p(F_X(q), F_X(r))^p = c \text{TV}(q, r) \quad \text{for all } q, r \in \Delta_M.$$

Then

$$\inf_{\widehat{W}} \sup_{q, r \in \Delta_M} \mathbb{E}(\widehat{W} - W_p(F_X(q), F_X(r)))^2 \lesssim_{d,p} N^{-1/p} = o(N^{-2/d}).$$

Thus exact deterministic one-scale identities cannot realize the supercritical difficulty of the unrestricted Euclidean problem.

Proof. By Theorem 4.1, the support is equilateral and has size at most $d + 1$. Apply Theorem 4.3 with $M \leq d + 1$. Since $d > 2p$, one has $1/p > 2/d$, so $N^{-1/p} = o(N^{-2/d})$. \square

5 The dyadic surrogate and the exact local theorem

This section proves Theorem 2.4. We begin with the canonical dyadic surrogate.

5.1 Dyadic partitions and the tree formula

For $j \geq 0$, let \mathcal{D}_j be the standard dyadic partition of $[0, 1]^d$ into 2^{jd} cubes of side length 2^{-j} . For distinct $x, y \in [0, 1]^d$, let

$$k(x, y) := \min\{j \geq 1 : x \text{ and } y \text{ lie in distinct cubes of } \mathcal{D}_j\},$$

and define the dyadic ultrametric

$$\rho(x, y) := 2^{-k(x, y)}, \quad \rho(x, x) := 0.$$

Write

$$W_{\rho,p}(P, Q) := \left(\inf_{\pi \in \Pi(P, Q)} \int \rho(x, y)^p d\pi(x, y) \right)^{1/p}, \quad T_\rho(P, Q) := W_{\rho,p}(P, Q)^p.$$

For a dyadic cube $R \in \mathcal{D}_j$, define

$$\Delta_R(P, Q) := P(R) - Q(R), \quad V_j(P, Q) := \sum_{R \in \mathcal{D}_j} |\Delta_R(P, Q)|.$$

Finally set

$$a_j := \frac{1}{2}(2^{-jp} - 2^{-(j+1)p}).$$

Proposition 5.1 (Exact dyadic tree formula). *For every $P, Q \in \mathcal{P}_d$,*

$$T_\rho(P, Q) = \sum_{j \geq 1} a_j V_j(P, Q).$$

Moreover, for every $J \geq 1$,

$$0 \leq T_\rho(P, Q) - \sum_{j=1}^J a_j V_j(P, Q) \leq 2 \sum_{j > J} a_j = 2^{-(J+1)p}.$$

Proof. The metric ρ^p is exactly the path metric on the complete rooted dyadic tree whose level- j edges have weight a_j : if two leaves first split at level k , then the path between them contains two copies of every edge weight a_j for $j \geq k$, so

$$2 \sum_{j \geq k} a_j = 2^{-kp} = \rho(x, y)^p.$$

For weighted tree metrics, the one-Wasserstein cost equals the sum, over tree edges, of edge length times the absolute mass imbalance across that edge. Grouping edges by level yields the displayed formula. The tail estimate follows from $0 \leq V_j(P, Q) \leq 2$. \square

Let X_J be the set of centers of the cubes in \mathcal{D}_J .

Proposition 5.2 (Comparison on the dyadic grid). *Let $J \geq 1$ and $x, y \in X_J$. Then*

$$\rho(x, y) \leq \|x - y\|_2 \leq 2\sqrt{d} \rho(x, y).$$

Consequently, for all $P, Q \in \mathcal{P}(X_J)$,

$$W_{\rho,p}(P, Q) \leq W_p(P, Q) \leq 2\sqrt{d} W_{\rho,p}(P, Q).$$

Proof. If $k = k(x, y)$, then some coordinate of $x - y$ differs by at least 2^{-k} , which gives the lower bound. Since x and y lie in the same level- $(k-1)$ dyadic cube, their Euclidean distance is at most its diameter $\sqrt{d}2^{-(k-1)} = 2\sqrt{d}2^{-k}$, giving the upper bound. The Wasserstein comparison follows by integrating the pointwise inequalities over arbitrary couplings and taking infima. \square

5.2 A direct estimator on the dyadic tree

We use the following discrete input of Jiao, Han, and Weissman [3].

Theorem 5.3 (Jiao–Han–Weissman). *There exists a universal constant $C_0 > 0$ such that the following holds. Let $S \geq 2$, let $N \geq 2$, and assume $\log N \leq C_0 \log S$. Then there exists an estimator \hat{L} based on N samples from each of two unknown probability vectors $r, s \in \Delta_S$ such that*

$$\sup_{r, s \in \Delta_S} \mathbb{E}(\hat{L} - \|r - s\|_1)^2 \leq C \frac{S}{N \log N}$$

for a universal constant $C < \infty$.

We also need the elementary empirical L_1 -bound.

Lemma 5.4 (Empirical L_1 -error on a finite alphabet). *Let r, s be distributions on an S -point alphabet, and let \hat{r}_N, \hat{s}_N be empirical histograms based on N samples each. Then*

$$\mathbb{E}(\|\hat{r}_N - \hat{s}_N\|_1 - \|r - s\|_1)^2 \leq \frac{4S}{N}.$$

Proof. By the triangle inequality,

$$\left| \|\hat{r}_N - \hat{s}_N\|_1 - \|r - s\|_1 \right| \leq \|\hat{r}_N - r\|_1 + \|\hat{s}_N - s\|_1.$$

The Cauchy–Schwarz inequality gives $\|\hat{r}_N - r\|_1^2 \leq S\|\hat{r}_N - r\|_2^2$, hence

$$\mathbb{E}\|\hat{r}_N - r\|_1^2 \leq S \sum_{i=1}^S \text{Var}(\hat{r}_{N,i}) \leq \frac{S}{N},$$

and the same for \hat{s}_N . □

Fix J_1 by

$$2^{J_1 d} \leq N \log N < 2^{(J_1+1)d}, \quad J_* := \left\lceil \frac{J_1}{2} \right\rceil.$$

For each $j \in \{1, \dots, J_1\}$, the level- j cell masses of P and Q form two probability vectors on an alphabet of size 2^{jd} . We estimate $V_j(P, Q)$ empirically for $j < J_*$ and by the JHW estimator for $J_* \leq j \leq J_1$. More precisely, let

$$\hat{V}_j^{\text{emp}} := V_j(P_N, Q_N), \quad 1 \leq j < J_*,$$

and for $J_* \leq j \leq J_1$, let \hat{V}_j^{JHW} be twice the JHW estimator of the L_1 -distance between the level- j cell-mass vectors, clipped to $[0, 2]$. Define

$$\hat{T}_\rho := \sum_{j=1}^{J_*-1} a_j \hat{V}_j^{\text{emp}} + \sum_{j=J_*}^{J_1} a_j \hat{V}_j^{\text{JHW}}, \quad \widehat{W}_\rho := \hat{T}_\rho^{1/p}.$$

Theorem 5.5 (Direct upper bound on the dyadic tree). *Assume $d > 2p$ and $n = m = N$. Then*

$$\sup_{P, Q \in \mathcal{P}_d} \mathbb{E}(\widehat{T}_\rho - T_\rho(P, Q))^2 \lesssim_{d,p} (N \log N)^{-2p/d},$$

and therefore

$$\sup_{P, Q \in \mathcal{P}_d} \mathbb{E}(\widehat{W}_\rho - W_{\rho,p}(P, Q))^2 \lesssim_{d,p} (N \log N)^{-2/d}.$$

Proof. Let Δ_j denote the estimation error at level j . By Theorem 5.1,

$$\widehat{T}_\rho - T_\rho(P, Q) = \sum_{j=1}^{J_1} a_j \Delta_j - \sum_{j>J_1} a_j V_j(P, Q).$$

Since $0 \leq V_j(P, Q) \leq 2$, Minkowski's inequality gives

$$(\mathbb{E}(\widehat{T}_\rho - T_\rho)^2)^{1/2} \leq \sum_{j=1}^{J_1} a_j (\mathbb{E}\Delta_j^2)^{1/2} + 2 \sum_{j>J_1} a_j.$$

For $j < J_*$, Theorem 5.4 yields

$$(\mathbb{E}\Delta_j^2)^{1/2} \leq 2\sqrt{2^{jd}/N}.$$

For $J_* \leq j \leq J_1$, the alphabet size is $S_j = 2^{jd}$, and because $j \geq J_* \geq J_1/2$, one has $\log N \leq C \log S_j$ for all large N . Hence Theorem 5.3 applies and gives

$$(\mathbb{E}\Delta_j^2)^{1/2} \lesssim \sqrt{2^{jd}/(N \log N)}.$$

The coarse sum is therefore bounded by

$$\sum_{j < J_*} 2^{-jp} \sqrt{2^{jd}/N} \lesssim_{d,p} N^{-1/2} 2^{J_*(d/2-p)} \lesssim_{d,p} 2^{-J_1 p},$$

and the fine sum by

$$\sum_{j=J_*}^{J_1} 2^{-jp} \sqrt{2^{jd}/(N \log N)} \lesssim_{d,p} \frac{2^{J_1(d/2-p)}}{\sqrt{N \log N}} \lesssim_{d,p} 2^{-J_1 p}.$$

The tail satisfies $\sum_{j > J_1} a_j \lesssim 2^{-J_1 p}$. Since $2^{J_1 d} \asymp N \log N$, we obtain

$$(\mathbb{E}(\widehat{T}_\rho - T_\rho)^2)^{1/2} \lesssim_{d,p} 2^{-J_1 p} \asymp (N \log N)^{-p/d}.$$

Finally,

$$|\widehat{W}_\rho - W_{\rho,p}|^2 \leq |\widehat{T}_\rho - T_\rho|^{2/p},$$

so monotonicity of L_r -norms yields the distance bound. \square

Theorem 5.6 (Exact dyadic-tree minimax law). *Assume $d > 2p$ and $n = m = N$. Then*

$$\inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}_d} \mathbb{E}(\widehat{W} - W_{\rho,p}(P, Q))^2 \asymp_{d,p} (N \log N)^{-2/d}.$$

Proof. The upper bound is Theorem 5.5. The matching lower bound is the specialization to the complete dyadic tree of the minimax lower theory of Wang, Cai, and Li [8] for transport distances on weighted trees. \square

Theorem 5.7 (Plug-in remains at $N^{-1/d}$ on the dyadic tree). *Assume $d > 2p$ and $n = m = N$. Then*

$$\sup_{P, Q \in \mathcal{P}_d} \mathbb{E}(W_{\rho,p}(P_N, Q_N) - W_{\rho,p}(P, Q))^2 \asymp_{d,p} N^{-2/d}.$$

Proof. For the upper bound, repeat the proof of Theorem 5.5 but use the empirical estimator at every level. This yields

$$\sup_{P, Q \in \mathcal{P}_d} \mathbb{E}(W_{\rho,p}(P_N, Q_N) - W_{\rho,p}(P, Q))^2 \lesssim_{d,p} N^{-2/d}.$$

For the lower bound, let J_0 satisfy $2^{J_0 d} \leq N < 2^{(J_0+1)d}$, and let $Q^{(0)}$ be the uniform distribution on the grid X_{J_0} . Then $W_{\rho,p}(Q^{(0)}, Q^{(0)}) = 0$, and the minimal nonzero ρ -distance among grid points is 2^{-J_0} . Hence for any empirical measures P_N, Q_N drawn from $Q^{(0)}$,

$$W_{\rho,p}(P_N, Q_N) \geq 2^{-J_0} \text{TV}(P_N, Q_N).$$

Since $|X_{J_0}| \asymp N$, Theorem 5.13 shows that $\mathbb{E} \text{TV}(P_N, Q_N) \gtrsim 1$, and therefore

$$\mathbb{E} W_{\rho,p}(P_N, Q_N)^2 \gtrsim_d 2^{-2J_0} \asymp N^{-2/d}.$$

□

Remark 5.8. The dyadic tree already exhibits the exact multiscale phenomenon one expects in Euclidean space: direct estimation attains the scale $(N \log N)^{-1/d}$, whereas plug-in remains at $N^{-1/d}$. The local Euclidean theorem proved below can be read as a transfer of this sharp dyadic picture to a full neighborhood of the actual diagonal.

5.3 Quantization and the local upper bound

Let $J = J(N)$ be the unique integer such that

$$2^{Jd} \leq N \log N < 2^{(J+1)d}.$$

Let $\kappa_J : [0, 1]^d \rightarrow X_J$ map each point to the center of the dyadic cube of level J that contains it, and write

$$P^{(J)} := (\kappa_J)_\# P.$$

Lemma 5.9 (Deterministic quantization bound). *For every $P \in \mathcal{P}_d$,*

$$W_p(P, P^{(J)}) \leq \sqrt{d} 2^{-J}.$$

Consequently, for every $P, Q \in \mathcal{P}_d$,

$$|W_p(P^{(J)}, Q^{(J)}) - W_p(P, Q)| \leq 2\sqrt{d} 2^{-J}.$$

Proof. Couple each point x with the center $\kappa_J(x)$ of its level- J cube. The diameter of that cube is $\sqrt{d} 2^{-J}$, which proves the first bound. The second follows by the triangle inequality. □

Theorem 5.10 (Local upper bound on the full Euclidean neighborhood). *Fix $A > 0$. There exists an estimator \widehat{W}_{loc} such that*

$$\sup_{(P,Q) \in \mathcal{L}_{A,N}} \mathbb{E} (\widehat{W}_{\text{loc}} - W_p(P, Q))^2 \lesssim_{A,d,p} \eta_N^2.$$

In particular,

$$\sup_{(P,Q) \in \mathcal{L}_{A,N}} \mathbb{E} |\widehat{W}_{\text{loc}} - W_p(P, Q)| \lesssim_{A,d,p} \eta_N.$$

Proof. Quantize the samples to level J and apply Theorem 5.5 to the resulting empirical measures supported on X_J . This yields an estimator \widehat{W}_{loc} satisfying

$$\sup_{P,Q \in \mathcal{P}_d} \mathbb{E} (\widehat{W}_{\text{loc}} - W_{\rho,p}(P^{(J)}, Q^{(J)}))^2 \lesssim_{d,p} \eta_N^2.$$

Now fix $(P, Q) \in \mathcal{L}_{A,N}$. By Theorem 5.9,

$$W_p(P^{(J)}, Q^{(J)}) \leq W_p(P, Q) + 2\sqrt{d} 2^{-J} \lesssim_{A,d} \eta_N.$$

Since $P^{(J)}$ and $Q^{(J)}$ are supported on the dyadic grid, Theorem 5.2 gives

$$0 \leq W_p(P^{(J)}, Q^{(J)}) - W_{\rho,p}(P^{(J)}, Q^{(J)}) \leq (2\sqrt{d} - 1)W_{\rho,p}(P^{(J)}, Q^{(J)}) \lesssim_{A,d} \eta_N.$$

Combining this with Theorem 5.9 yields

$$|W_p(P, Q) - W_{\rho,p}(P^{(J)}, Q^{(J)})| \lesssim_{A,d} \eta_N.$$

Therefore

$$|\widehat{W}_{\text{loc}} - W_p(P, Q)| \leq |\widehat{W}_{\text{loc}} - W_{\rho,p}(P^{(J)}, Q^{(J)})| + C_{A,d}\eta_N,$$

which proves the squared bound. The absolute-risk bound follows from Jensen's inequality. \square

5.4 Local lower bound from the Niles–Weed–Rigollet construction

The next result extracts a stronger conclusion from the lower-bound proof of Niles–Weed and Rigollet [7]: their hard instances already lie inside a shrinking Euclidean neighborhood of the diagonal.

Theorem 5.11 (Local lower bound). *Assume $d > 2p$. There exist constants $A_0 = A_0(d, p)$, $c_0 = c_0(d, p) > 0$, and $N_0 = N_0(d, p)$ such that for all $N \geq N_0$ and all estimators \widehat{W} ,*

$$\sup_{(P,Q) \in \mathcal{L}_{A_0,N}} \mathbb{P}\left(|\widehat{W} - W_p(P, Q)| \geq c_0\eta_N\right) \geq c_0.$$

Consequently,

$$\begin{aligned} \inf_{\widehat{W}} \sup_{(P,Q) \in \mathcal{L}_{A_0,N}} \mathbb{E}|\widehat{W} - W_p(P, Q)| &\gtrsim_{d,p} \eta_N, \\ \inf_{\widehat{W}} \sup_{(P,Q) \in \mathcal{L}_{A_0,N}} \mathbb{E}(\widehat{W} - W_p(P, Q))^2 &\gtrsim_{d,p} \eta_N^2. \end{aligned}$$

Proof. We revisit the proof of [7, Theorem 11], keeping track of the size of the hard instances. Let u be the uniform distribution on $[m]$, where $m \asymp N \log N$ will be chosen below. Let D_m , $D_{m,\delta}^-$, and D_m^+ be the composite hypotheses from [7, Proposition 10]:

$$\begin{aligned} D_m &:= \{q : \chi^2(q, u) \leq 9\}, \\ D_{m,\delta}^- &:= \{q \in D_m : \text{TV}(q, u) \leq \delta\}, \\ D_m^+ &:= \{q \in D_m : \text{TV}(q, u) \geq 1/4\}. \end{aligned}$$

Let $F : [m] \rightarrow [0, 1]^d$ be the random map from [7, Proposition 9], and define

$$\mu_{F,q} := F_{\#}q, \quad \nu_F := F_{\#}u.$$

With $m \asymp N \log N$, the proof of [7, Theorem 11] yields constants $A_0, c_0 > 0$ and $\Delta_N \asymp \eta_N$ such that, with probability at least 0.9 over F ,

$$W_p(\mu_{F,q}, \nu_F) \leq A_0\eta_N \quad (q \in D_m),$$

while the “null” and “alternative” subfamilies satisfy

$$W_p(\mu_{F,q}, \nu_F) \leq \Delta_N \quad (q \in D_{m,\delta}^-),$$

$$W_p(\mu_{F,q}, \nu_F) \geq 3\Delta_N \quad (q \in D_m^+),$$

on the same good event. In particular, all hard pairs belong to $\mathcal{L}_{A_0, N}$.

Now apply the same threshold test as in [7]:

$$\psi := \mathbf{1}\{\widehat{W} \leq 2\Delta_N\}.$$

On the null family, the event $\{\psi = 0\}$ implies an estimation error of at least Δ_N ; on the alternative family, the event $\{\psi = 1\}$ implies the same. By [7, Proposition 10], the sum of the maximal testing errors on the two composite hypotheses is bounded below by a positive constant. Averaging over F and fixing one realization for which the same lower bound holds gives

$$\sup_{(P,Q) \in \mathcal{L}_{A_0, N}} \mathbb{P}(|\widehat{W} - W_p(P, Q)| \geq \Delta_N) \geq c_0.$$

Since $\Delta_N \asymp \eta_N$, the probability bound follows after absorbing constants. The risk lower bounds are immediate from $\mathbb{E}|Z| \geq a\mathbb{P}(|Z| \geq a)$ and $\mathbb{E}Z^2 \geq a^2\mathbb{P}(|Z| \geq a)$. \square

Corollary 5.12 (Exact local minimax law). *Assume $d > 2p$ and $n = m = N$. There exists $A_0 = A_0(d, p)$ such that*

$$\inf_{\widehat{W}} \sup_{(P,Q) \in \mathcal{L}_{A_0, N}} \mathbb{E}|\widehat{W} - W_p(P, Q)| \asymp_{d,p} \eta_N,$$

$$\inf_{\widehat{W}} \sup_{(P,Q) \in \mathcal{L}_{A_0, N}} \mathbb{E}(\widehat{W} - W_p(P, Q))^2 \asymp_{d,p} \eta_N^2.$$

Proof. Combine Theorems 5.10 and 5.11. \square

5.5 The empirical plug-in estimator is still locally suboptimal

The local exact law concerns the minimax benchmark. We now show that the empirical plug-in estimator remains at the slower $N^{-1/d}$ scale even on $\mathcal{L}_{A, N}$.

Lemma 5.13 (Empirical total variation at the uniform law). *Let U_M be the uniform distribution on an M -point alphabet, and let $\widehat{U}_N, \widehat{U}'_N$ be two independent empirical measures based on N samples from U_M . Then*

$$\mathbb{E} \text{TV}(\widehat{U}_N, \widehat{U}'_N) \geq \left(1 - \frac{1}{M}\right)^{2N-1}.$$

In particular, if $M \geq cN$, then $\mathbb{E} \text{TV}(\widehat{U}_N, \widehat{U}'_N) \geq c' > 0$, where c' depends only on c .

Proof. Let $B, B' \stackrel{\text{i.i.d.}}{\sim} \text{Bin}(N, 1/M)$. By symmetry of the M symbols,

$$\mathbb{E} \text{TV}(\widehat{U}_N, \widehat{U}'_N) = \frac{M}{2N} \mathbb{E}|B - B'|.$$

Now

$$\mathbb{E}|B - B'| \geq 2\mathbb{P}(B = 1)\mathbb{P}(B' = 0) = 2\frac{N}{M} \left(1 - \frac{1}{M}\right)^{2N-1},$$

which gives the claim. \square

Theorem 5.14 (Local plug-in lower bound). *Assume $d > 2p$ and $n = m = N$. For every fixed $A > 0$,*

$$\sup_{(P,Q) \in \mathcal{L}_{A,N}} \mathbb{E}(W_p(P_N, Q_N) - W_p(P, Q))^2 \gtrsim_{d,p} N^{-2/d}.$$

Proof. Choose $M = N$ points $x_1, \dots, x_M \in [0, 1]^d$ with pairwise Euclidean separation at least $c_d N^{-1/d}$, and let

$$Q_0 := \frac{1}{M} \sum_{i=1}^M \delta_{x_i}.$$

Then $W_p(Q_0, Q_0) = 0$, so $(Q_0, Q_0) \in \mathcal{L}_{A,N}$ for every $A > 0$. If P_N and Q_N are two independent empirical measures based on Q_0 , any nonzero move between support points costs at least $c_d N^{-1/d}$, hence

$$W_p(P_N, Q_N) \geq c_d N^{-1/d} \text{TV}(P_N, Q_N).$$

By Theorem 5.13, $\mathbb{E} \text{TV}(P_N, Q_N) \geq c > 0$, so

$$\mathbb{E} W_p(P_N, Q_N) \gtrsim_d N^{-1/d}.$$

Since the target value is zero at (Q_0, Q_0) , Jensen's inequality gives

$$\mathbb{E}(W_p(P_N, Q_N) - W_p(Q_0, Q_0))^2 \geq (\mathbb{E} W_p(P_N, Q_N))^2 \gtrsim_d N^{-2/d}.$$

□

Corollary 5.15 (Full local logarithmic separation). *Assume $d > 2p$ and $n = m = N$. There exists $A_0 = A_0(d, p)$ such that*

$$\inf_{\widehat{W}} \sup_{(P,Q) \in \mathcal{L}_{A_0,N}} \mathbb{E}(\widehat{W} - W_p(P, Q))^2 \asymp_{d,p} (N \log N)^{-2/d},$$

whereas plug-in satisfies

$$\sup_{(P,Q) \in \mathcal{L}_{A_0,N}} \mathbb{E}(W_p(P_N, Q_N) - W_p(P, Q))^2 \gtrsim_{d,p} N^{-2/d}.$$

Thus direct estimation beats plug-in by a genuine logarithmic factor even on the full Euclidean neighborhood of the diagonal.

Proof. Combine Theorems 5.12 and 5.14. □

6 A critical-grid reduction of the unrestricted problem

The exact local theorem shows that the scale $(N \log N)^{-1/d}$ is already intrinsic in Euclidean space. The next result identifies the corresponding finite-state model.

Fix $L \geq 1$, and let $J = J(N, L)$ be the unique integer such that

$$2^{Jd} \leq LN \log N < 2^{(J+1)d}.$$

Let X_J be the set of centers of the cubes in \mathcal{D}_J . Define

$$\mathfrak{G}_{N,L,d,p}^{\text{abs}} := \inf_{\widehat{W}} \sup_{\mu, \nu \in \mathcal{P}(X_J)} \mathbb{E} |\widehat{W} - W_p(\mu, \nu)|,$$

$$\mathfrak{G}_{N,L,d,p}^{\text{sq}} := \inf_{\widehat{W}} \sup_{\mu, \nu \in \mathcal{P}(X_J)} \mathbb{E}(\widehat{W} - W_p(\mu, \nu))^2.$$

Theorem 6.1 (Critical-grid reduction). *Under the assumptions $d > 2p$ and $n = m = N$,*

$$\mathfrak{G}_{N,L,d,p}^{\text{abs}} \leq \mathfrak{M}_{N,N,d,p}^{\text{abs}} \leq \mathfrak{G}_{N,L,d,p}^{\text{abs}} + 2\sqrt{d}2^{-J},$$

$$\mathfrak{G}_{N,L,d,p}^{\text{sq}} \leq \mathfrak{M}_{N,N,d,p}^{\text{sq}} \leq 2\mathfrak{G}_{N,L,d,p}^{\text{sq}} + 8d2^{-2J}.$$

Consequently, for every fixed sufficiently large L ,

$$\mathfrak{G}_{N,L,d,p}^{\text{abs}} \asymp_{L,d,p} (N \log N)^{-1/d} \iff \mathfrak{M}_{N,N,d,p}^{\text{abs}} \asymp_{d,p} (N \log N)^{-1/d},$$

with the analogous equivalence for squared risk.

Proof. The lower inequalities are immediate because $\mathcal{P}(X_J) \subset \mathcal{P}_d$. For the upper bounds, let \widehat{W}_J be any estimator for the grid problem. Given arbitrary $P, Q \in \mathcal{P}_d$, quantize the samples to level J and feed the resulting observations to \widehat{W}_J . This produces an estimator \widetilde{W} whose target is $W_p(P^{(J)}, Q^{(J)})$. By Theorem 5.9,

$$|W_p(P^{(J)}, Q^{(J)}) - W_p(P, Q)| \leq 2\sqrt{d}2^{-J}.$$

The absolute-risk bound follows immediately. For squared loss, use $(a + b)^2 \leq 2a^2 + 2b^2$. The equivalence follows because $2^{-J} \asymp L^{-1/d}(N \log N)^{-1/d}$, while Theorem 5.12 provides the matching lower bound for the unrestricted problem. \square

Remark 6.2. The critical alphabet size is neither fixed nor of order N , but of order $N \log N$. This is exactly the sample-size enlargement scale at which the classical discrete L_1 -distance problem changes from the empirical rate $\sqrt{S/N}$ to the direct-estimation rate $\sqrt{S/(N \log N)}$; see [3].

7 Separated classes: the smooth off-diagonal regime

The exact local theorem controls the diagonal neighborhood. At the opposite extreme, plug-in already attains the target scale as soon as the true distance is sufficiently large. This leaves only a middle annulus in which the genuinely mixed-scale geometry must be analyzed.

Throughout this section assume $d > 2p$, $n = m = N$, and set

$$\eta_N := (N \log N)^{-1/d}, \quad \beta_p := p \wedge 2.$$

For $r > 0$, define the separated class

$$\mathcal{S}_r := \{(P, Q) \in \mathcal{P}_d^2 : W_p(P, Q) \geq r\}.$$

Proposition 7.1 (Smooth-cost upper bound for the empirical p -cost). *For every $p \geq 1$,*

$$\sup_{P, Q \in \mathcal{P}_d} \mathbb{E}|W_p(P_N, Q_N)^p - W_p(P, Q)^p| \lesssim_{d,p} N^{-\beta_p/d}.$$

Proof. This is exactly [6, Corollary 3(i)] specialized to the Euclidean cost $c(x, y) = \|x - y\|_2^p$ on the compact cube $[0, 1]^d$. \square

Lemma 7.2 (Power-to-distance conversion off the diagonal). *Let $p \geq 1$, $r > 0$, $y \geq r$, and $x \geq 0$. Then*

$$|x - y| \leq r^{1-p} |x^p - y^p|.$$

Proof. If $x \geq y$, then

$$x^p - y^p = p \int_y^x t^{p-1} dt \geq pr^{p-1}(x - y) \geq r^{p-1}(x - y).$$

If $0 \leq x \leq y$, consider

$$g(x) := y^p - x^p - r^{p-1}(y - x).$$

The function g is concave on $[0, y]$ because $g''(x) = -p(p-1)x^{p-2} \leq 0$, and $g(0) = y(y^{p-1} - r^{p-1}) \geq 0$, $g(y) = 0$. Hence $g(x) \geq 0$ for all $x \in [0, y]$, which is equivalent to the desired bound. \square

Theorem 7.3 (Separated-class upper bound in absolute loss). *For every $r > 0$,*

$$\sup_{(P,Q) \in \mathcal{S}_r} \mathbb{E} |W_p(P_N, Q_N) - W_p(P, Q)| \lesssim_{d,p} r^{1-p} N^{-\beta_p/d}.$$

Proof. Fix $(P, Q) \in \mathcal{S}_r$ and set $x := W_p(P_N, Q_N)$, $y := W_p(P, Q)$. Since $y \geq r$, Theorem 7.2 gives

$$|W_p(P_N, Q_N) - W_p(P, Q)| \leq r^{1-p} |W_p(P_N, Q_N)^p - W_p(P, Q)^p|.$$

Taking expectations and applying Theorem 7.1 proves the claim. \square

For $p > 1$, define the threshold

$$\tau_{N,p} := (N^{-\beta_p/d} \eta_N^{-1})^{1/(p-1)}.$$

Explicitly,

$$\tau_{N,p} = \begin{cases} N^{-1/d} (\log N)^{1/((p-1)d)}, & 1 < p < 2, \\ N^{-1/d} (\log N)^{1/d}, & p = 2, \\ N^{-1/((p-1)d)} (\log N)^{1/((p-1)d)}, & p > 2. \end{cases}$$

Corollary 7.4 (The target scale is already achieved on sufficiently separated classes). *Assume $p > 1$. There exists $B = B(d, p) \geq 1$ such that*

$$\sup_{(P,Q) \in \mathcal{S}_{B\tau_{N,p}}} \mathbb{E} |W_p(P_N, Q_N) - W_p(P, Q)| \lesssim_{d,p} \eta_N.$$

Proof. Apply Theorem 7.3 with $r = B\tau_{N,p}$. By the definition of $\tau_{N,p}$,

$$(B\tau_{N,p})^{1-p} N^{-\beta_p/d} = B^{1-p} \eta_N.$$

Choose B large enough to absorb the implicit constant. \square

Remark 7.5. Together, Theorems 5.12 and 7.4 show that the target scale η_N is already attained on the full diagonal neighborhood and on all sufficiently separated classes. The only remaining distances at this stage lie in an explicit middle annulus, whose geometry is resolved in the next section by the mixed-scale two-ball analysis.

8 Positive-distance neighborhoods and mixed-scale geometry

This section proves Theorems 2.6 and 2.7. The first theorem gives a coarse positive-distance picture, while the second resolves the full mixed-scale geometry on two-ball product classes.

8.1 Separated product neighborhoods

Recall the fixed template sets

$$X_\star := \{u \in \mathbb{R}^d : \|u\|_2 \leq \gamma\}, \quad Y_\star := e_1 + X_\star, \quad \gamma := 1/20,$$

and define for $0 < r \leq 1/4$

$$X_r := z_\star + rX_\star, \quad Y_r := z_\star + rY_\star, \quad \mathcal{C}_r := \mathcal{P}(X_r) \times \mathcal{P}(Y_r),$$

where $z_\star = (1/4, 1/2, \dots, 1/2)$.

Lemma 8.1 (Template geometry). *For every $0 < r \leq 1/4$ and every $x \in X_r, y \in Y_r$,*

$$\frac{9}{10}r \leq \|x - y\|_2 \leq \frac{11}{10}r.$$

Consequently, for every $(P, Q) \in \mathcal{C}_r$,

$$\frac{9}{10}r \leq W_p(P, Q) \leq \frac{11}{10}r.$$

Proof. Write $x = z_\star + ru$ and $y = z_\star + rv$ with $u \in X_\star$ and $v \in Y_\star = e_1 + X_\star$. Then $u - v \in -e_1 + \{w : \|w\|_2 \leq 2\gamma\}$, so $1 - 2\gamma \leq \|u - v\|_2 \leq 1 + 2\gamma$. Multiplying by r gives the pointwise bound; integrating over couplings yields the statement for W_p . \square

Let

$$S_r(u) := z_\star + ru, \quad \bar{P} := (S_r^{-1})_\# P \in \mathcal{P}(X_\star), \quad \bar{Q} := (S_r^{-1})_\# Q \in \mathcal{P}(Y_\star).$$

Since the transport cost scales exactly,

$$W_p(P, Q)^p = r^p T_\star(\bar{P}, \bar{Q}), \quad T_\star(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int \|u - v\|_2^p d\pi(u, v).$$

Theorem 8.2 (Full product neighborhoods away from the diagonal). *Assume $d \geq 5$ and $p > 1$. There exist constants $0 < c_{d,p} \leq C_{d,p} < \infty$ such that for every $0 < r \leq 1/4$,*

$$c_{d,p} r^p (N \log N)^{-2/d} \leq \inf_{\hat{T}} \sup_{(P, Q) \in \mathcal{C}_r} \mathbb{E} |\hat{T} - W_p(P, Q)^p| \leq C_{d,p} r^p N^{-2/d}.$$

Consequently,

$$c_{d,p} r (N \log N)^{-2/d} \leq \inf_{\hat{W}} \sup_{(P, Q) \in \mathcal{C}_r} \mathbb{E} |\hat{W} - W_p(P, Q)| \leq C_{d,p} r N^{-2/d}.$$

Proof. On $X_\star \times Y_\star$, the cost $c_\star(u, v) = \|u - v\|_2^p$ is smooth and uniformly strongly convex in the sense required by Manole and Niles-Weed [6]: the difference variable stays in a compact annulus away from the origin, so the Hessian of $z \mapsto \|z\|_2^p$ is uniformly bounded above and below. Hence [6, Theorem 22] yields the lower bound $(N \log N)^{-2/d}$ for the template problem,

while [6, Corollary 3(ii)] yields the matching plug-in upper bound $N^{-2/d}$. Scaling back by r^p gives the cost statement. Finally, by Theorem 8.1, the map $x \mapsto x^p$ is bi-Lipschitz on the interval $[\frac{9}{10}r, \frac{11}{10}r]$, so the distance statement follows by converting between W_p and W_p^p . \square

8.2 Exact first-plus-second-order decomposition on two-ball classes

Fix

$$B := \{u \in \mathbb{R}^d : \|u\|_2 \leq 1\}, \quad z_\diamond := \left(\frac{1}{4}, \frac{1}{2}, \dots, \frac{1}{2}\right),$$

and parameters

$$0 < r \leq 1/4, \quad 0 < s \leq r/20, \quad \kappa := s/r \in (0, 1/20].$$

Define

$$X_{r,s} := z_\diamond + sB, \quad Y_{r,s} := z_\diamond + re_1 + sB, \quad \mathcal{C}_{r,s} := \mathcal{P}(X_{r,s}) \times \mathcal{P}(Y_{r,s}).$$

For $(P, Q) \in \mathcal{C}_{r,s}$, set

$$\begin{aligned} S_{r,s}(u) &:= z_\diamond + su, & T_{r,s}(v) &:= z_\diamond + re_1 + sv, \\ \mu &:= (S_{r,s}^{-1})_{\#}P \in \mathcal{P}(B), & \nu &:= (T_{r,s}^{-1})_{\#}Q \in \mathcal{P}(B). \end{aligned}$$

Let

$$h_p(z) := \|z\|_2^p, \quad b_p := \nabla h_p(-e_1) = -pe_1.$$

For $t \in 2B$, define the normalized second-order remainder

$$g_{\kappa,p}(t) := \frac{h_p(-e_1 + \kappa t) - h_p(-e_1) - \kappa b_p \cdot t}{\kappa^2},$$

and let

$$U_{\kappa,p}(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int g_{\kappa,p}(u - v) d\pi(u, v).$$

Lemma 8.3 (Two-ball geometry). *For every $(P, Q) \in \mathcal{C}_{r,s}$,*

$$\frac{9}{10}r \leq W_p(P, Q) \leq \frac{11}{10}r.$$

Proof. For $x = z_\diamond + su$ and $y = z_\diamond + re_1 + sv$ with $u, v \in B$, one has

$$x - y = -re_1 + s(u - v),$$

so $r - 2s \leq \|x - y\|_2 \leq r + 2s$. Because $s \leq r/20$, this implies the stated bounds, which then pass to W_p by integration over couplings. \square

Proposition 8.4 (Exact decomposition on $\mathcal{C}_{r,s}$). *For every $(P, Q) \in \mathcal{C}_{r,s}$,*

$$W_p(P, Q)^p = r^p + r^{p-1}s b_p \cdot \left(\int u d\mu(u) - \int v d\nu(v) \right) + r^{p-2}s^2 U_{\kappa,p}(\mu, \nu).$$

Proof. If $\pi \in \Pi(\mu, \nu)$, then $(S_{r,s}, T_{r,s})_{\#}\pi \in \Pi(P, Q)$, and for $(u, v) \in B \times B$,

$$S_{r,s}(u) - T_{r,s}(v) = r(-e_1 + \kappa(u - v)).$$

Therefore

$$\begin{aligned} \int \|x - y\|_2^p d(S_{r,s}, T_{r,s})_{\#} \pi(x, y) &= r^p \int h_p(-e_1 + \kappa(u - v)) d\pi(u, v) \\ &= r^p + r^p \kappa b_p \cdot \int (u - v) d\pi(u, v) \\ &\quad + r^p \kappa^2 \int g_{\kappa,p}(u - v) d\pi(u, v). \end{aligned}$$

The constant and linear terms are independent of π , so taking the infimum over couplings proves the identity. \square

The residual term is controlled by a fixed quadratic transport cost. Let

$$H_p := \nabla^2 h_p(-e_1) = pI_d + p(p-2)e_1 e_1^\top, \quad q_p^\circ(t) := \frac{1}{2} t^\top H_p t.$$

Then

$$q_p^\circ(t) = \frac{p}{2} \sum_{i=2}^d t_i^2 + \frac{p(p-1)}{2} t_1^2.$$

Lemma 8.5 (Quadratic comparison for the residual kernel). *There exists $\kappa_0 = \kappa_0(d, p) \in (0, 1/20]$ such that for every $0 < \kappa \leq \kappa_0$ and every $t \in 2B$,*

$$\frac{1}{2} q_p^\circ(t) \leq g_{\kappa,p}(t) \leq \frac{3}{2} q_p^\circ(t).$$

Consequently, for all $\mu, \nu \in \mathcal{P}(B)$,

$$\frac{1}{2} T_{q_p^\circ}(\mu, \nu) \leq U_{\kappa,p}(\mu, \nu) \leq \frac{3}{2} T_{q_p^\circ}(\mu, \nu),$$

where

$$T_{q_p^\circ}(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int q_p^\circ(u - v) d\pi(u, v).$$

Proof. By Taylor's theorem with integral remainder,

$$g_{\kappa,p}(t) = \int_0^1 (1 - \tau) t^\top \nabla^2 h_p(-e_1 + \tau \kappa t) t d\tau.$$

For $t \in 2B$ and $0 < \kappa \leq 1/20$, all points $-e_1 + \tau \kappa t$ stay in a fixed compact annulus about $-e_1$. Since $\nabla^2 h_p$ is continuous there and $H_p = \nabla^2 h_p(-e_1)$ is positive definite, shrinking κ_0 if necessary gives

$$\frac{1}{2} H_p \preceq \nabla^2 h_p(-e_1 + \tau \kappa t) \preceq \frac{3}{2} H_p.$$

Integrating against $(1 - \tau) d\tau$ yields the pointwise bound, and the transport-cost comparison follows by integration over couplings and infimization. \square

The next proposition transfers the exact local diagonal theorem, already proved in Section 5, to the quadratic reference cost $T_{q_p^\circ}$.

Proposition 8.6 (Local quadratic lower bound). *There exist constants $a_{d,p}, c_{d,p} > 0$ such that for all sufficiently large N ,*

$$\inf_{\widehat{U}} \sup_{\substack{\mu, \nu \in \mathcal{P}(B): \\ T_{q_p^\circ}(\mu, \nu) \leq a_{d,p} \eta_N^2}} \mathbb{E} |\widehat{U} - T_{q_p^\circ}(\mu, \nu)| \geq c_{d,p} \eta_N^2.$$

Proof. There exists a positive-definite matrix A_p such that $q_p^\circ(t) = \|A_p t\|_2^2$. Hence

$$T_{q_p^\circ}(\mu, \nu)^{1/2} = W_2((A_p)_\# \mu, (A_p)_\# \nu).$$

The image $A_p(B)$ is a fixed ellipsoid. After a fixed affine change of variables, estimating this W_2 -distance on $A_p(B)$ is equivalent, up to constants depending only on d and p , to estimating the standard Euclidean W_2 -distance on a compact subset of $[0, 1]^d$. Because $d \geq 5 > 4$, Theorem 5.12 with $p = 2$ yields a lower bound of order η_N for the distance and hence of order η_N^2 for its square. Finally,

$$|\sqrt{a} - \sqrt{b}|^2 \leq |a - b| \quad (a, b \geq 0),$$

so the absolute risk for the quadratic cost dominates the squared risk for its square-root distance. \square

Proposition 8.7 (Residual lower bound on the local quadratic scale). *There exist constants $\kappa_1 = \kappa_1(d, p) \in (0, \kappa_0]$, $a_{d,p}^\# > 0$, and $c_{d,p}^\# > 0$ such that for every $0 < \kappa \leq \kappa_1$ and all sufficiently large N ,*

$$\inf_{\widehat{U}} \sup_{\substack{\mu, \nu \in \mathcal{P}(B): \\ T_{q_p^\circ}(\mu, \nu) \leq a_{d,p}^\# \eta_N^2}} \mathbb{E} |\widehat{U} - U_{\kappa,p}(\mu, \nu)| \geq c_{d,p}^\# \eta_N^2.$$

Proof. On the class

$$\mathcal{G}_N := \{(\mu, \nu) : T_{q_p^\circ}(\mu, \nu) \leq a_{d,p}^\# \eta_N^2\},$$

Theorem 8.5 gives

$$|U_{\kappa,p}(\mu, \nu) - T_{q_p^\circ}(\mu, \nu)| \leq \frac{1}{2} T_{q_p^\circ}(\mu, \nu) \leq \frac{1}{2} a_{d,p}^\# \eta_N^2.$$

Hence any estimator of $U_{\kappa,p}$ induces an estimator of $T_{q_p^\circ}$ with an additive error at most $\frac{1}{2} a_{d,p}^\# \eta_N^2$. Invoking Theorem 8.6 and shrinking $a_{d,p}$ if necessary yields the claim. \square

8.3 The mixed-scale phase diagram

The first-order term in Theorem 8.4 already yields a parametric lower bound.

Lemma 8.8 (Parametric lower bound on $\mathcal{C}_{r,s}$). *There exists $c_{d,p} > 0$ such that*

$$\inf_{\widehat{T}} \sup_{(P,Q) \in \mathcal{C}_{r,s}} \mathbb{E} |\widehat{T} - W_p(P, Q)^p| \geq c_{d,p} r^{p-1} s N^{-1/2},$$

$$\inf_{\widehat{W}} \sup_{(P,Q) \in \mathcal{C}_{r,s}} \mathbb{E} |\widehat{W} - W_p(P, Q)| \geq c_{d,p} s N^{-1/2}.$$

Proof. Restrict to the one-parameter subfamily

$$P_t = t \delta_{z_\circ + s e_1} + (1-t) \delta_{z_\circ - s e_1}, \quad Q = \delta_{z_\circ + r e_1}, \quad t \in [1/4, 3/4].$$

Then $(P_t, Q) \in \mathcal{C}_{r,s}$, and

$$W_p(P_t, Q)^p = t(r-s)^p + (1-t)(r+s)^p = \alpha_{r,s,p} + \beta_{r,s,p} t,$$

with $|\beta_{r,s,p}| \asymp_p r^{p-1}s$ by the mean value theorem. Estimating $W_p(P_t, Q)^p$ is therefore equivalent, up to constants, to estimating a Bernoulli mean, which has minimax absolute risk $\asymp N^{-1/2}$. Since $x \mapsto x^{1/p}$ is bi-Lipschitz on $[(9/10)r^p, (11/10)r^p]$, the same argument converts the lower bound to the distance itself. \square

We now construct the corresponding upper bound by estimating the linear term directly and the residual term by plug-in for the smooth bounded cost $g_{\kappa,p}$.

Proposition 8.9 (Direct estimator on $\mathcal{C}_{r,s}$). *There exists an estimator $\widehat{T}_{r,s}$ such that*

$$\sup_{(P,Q) \in \mathcal{C}_{r,s}} \mathbb{E} |\widehat{T}_{r,s} - W_p(P, Q)^p| \lesssim_{d,p} r^{p-1} s N^{-1/2} + r^{p-2} s^2 N^{-2/d}.$$

Consequently, there exists an estimator $\widehat{W}_{r,s}$ such that

$$\sup_{(P,Q) \in \mathcal{C}_{r,s}} \mathbb{E} |\widehat{W}_{r,s} - W_p(P, Q)| \lesssim_{d,p} s N^{-1/2} + \frac{s^2}{r} N^{-2/d}.$$

Proof. Let $U_i = S_{r,s}^{-1}(X_i) \in B$ and $V_i = T_{r,s}^{-1}(Y_i) \in B$, and write

$$\widehat{\mu}_N := \frac{1}{N} \sum_{i=1}^N \delta_{U_i}, \quad \widehat{\nu}_N := \frac{1}{N} \sum_{i=1}^N \delta_{V_i}, \quad \overline{U}_N := \frac{1}{N} \sum_{i=1}^N U_i, \quad \overline{V}_N := \frac{1}{N} \sum_{i=1}^N V_i.$$

Define

$$\widehat{U}_{\kappa,p} := T_{g_{\kappa,p}}(\widehat{\mu}_N, \widehat{\nu}_N)$$

and

$$\widehat{T}_{r,s} := r^p + r^{p-1} s b_p \cdot (\overline{U}_N - \overline{V}_N) + r^{p-2} s^2 \widehat{U}_{\kappa,p}.$$

By Theorem 8.4,

$$\begin{aligned} \widehat{T}_{r,s} - W_p(P, Q)^p &= r^{p-1} s b_p \cdot \left[(\overline{U}_N - \overline{V}_N) - \left(\int u d\mu - \int v d\nu \right) \right] \\ &\quad + r^{p-2} s^2 (\widehat{U}_{\kappa,p} - U_{\kappa,p}(\mu, \nu)). \end{aligned}$$

The first term is bounded by $\lesssim_{d,p} r^{p-1} s N^{-1/2}$, because the coordinates of U_i and V_i are uniformly bounded on B . For the second term, the cost $g_{\kappa,p}$ is a bounded C^2 function on $2B$ with a C^2 -norm uniformly bounded over $0 < \kappa \leq 1/20$, since all arguments $-e_1 + \kappa t$ remain in a fixed annulus away from the origin. The upper-bound part of Manole and Niles-Weed [6, Theorem 2] therefore gives

$$\sup_{\mu, \nu \in \mathcal{P}(B)} \mathbb{E} |\widehat{U}_{\kappa,p} - U_{\kappa,p}(\mu, \nu)| \lesssim_{d,p} N^{-2/d}.$$

This proves the cost bound. The distance bound follows from Theorem 8.3: after clipping $\widehat{T}_{r,s}$ to the interval $[(9/10)r^p, (11/10)r^p]$, the map $x \mapsto x^{1/p}$ is Lipschitz there with constant $\lesssim_p r^{1-p}$. \square

Theorem 8.10 (Mixed-scale lower envelope on the full two-ball class). *There exist $\kappa_\star = \kappa_\star(d, p) \in (0, \kappa_1]$ and $c_{d,p} > 0$ such that for every $0 < r \leq 1/4$ and every $0 < s \leq \kappa_\star r$,*

$$\inf_{\widehat{T}} \sup_{(P,Q) \in \mathcal{C}_{r,s}} \mathbb{E} |\widehat{T} - W_p(P, Q)^p| \geq c_{d,p} \max \left\{ r^{p-1} s N^{-1/2}, r^{p-2} s^2 \eta_N^2 \right\},$$

$$\inf_{\widehat{W}} \sup_{(P,Q) \in \mathcal{C}_{r,s}} \mathbb{E} |\widehat{W} - W_p(P,Q)| \geq c_{d,p} \max \left\{ sN^{-1/2}, \frac{s^2}{r} \eta_N^2 \right\}.$$

Proof. The parametric part is exactly Theorem 8.8. For the residual part, restrict attention to those pairs $(P, Q) \in \mathcal{C}_{r,s}$ whose rescaled marginals satisfy

$$T_{q_p^\circ}(\mu, \nu) \leq a_{d,p}^\# \eta_N^2.$$

Now reveal to the estimator the exact value of the first-order term

$$L(P, Q) := r^p + r^{p-1} s b_p \cdot \left(\int u d\mu - \int v d\nu \right).$$

This can only make the problem easier. By Theorem 8.4, estimating $W_p(P, Q)^p$ is then equivalent to estimating $r^{p-2} s^2 U_{\kappa,p}(\mu, \nu)$. Applying Theorem 8.7 gives

$$\inf_{\widehat{T}} \sup_{(P,Q) \in \mathcal{C}_{r,s}} \mathbb{E} |\widehat{T} - W_p(P, Q)^p| \gtrsim_{d,p} r^{p-2} s^2 \eta_N^2.$$

Taking the maximum with the parametric lower bound yields the cost statement. For the distance itself, use Theorem 8.3: on $[\frac{9}{10}r, \frac{11}{10}r]$, the map $x \mapsto x^p$ is Lipschitz with constant $\lesssim_p r^{p-1}$, so a lower bound of size $r^{p-2} s^2 \eta_N^2$ for the cost translates into a lower bound of size $(s^2/r) \eta_N^2$ for the distance. \square

Corollary 8.11 (Phase diagram on $\mathcal{C}_{r,s}$). *There exist $\kappa_\star = \kappa_\star(d, p) \in (0, 1/20]$ and constants $0 < c_{d,p} \leq C_{d,p} < \infty$ such that for every $0 < r \leq 1/4$ and $0 < s \leq \kappa_\star r$,*

$$\begin{aligned} c_{d,p} \max \left\{ sN^{-1/2}, \frac{s^2}{r} \eta_N^2 \right\} &\leq \inf_{\widehat{W}} \sup_{(P,Q) \in \mathcal{C}_{r,s}} \mathbb{E} |\widehat{W} - W_p(P, Q)| \\ &\leq C_{d,p} \left(sN^{-1/2} + \frac{s^2}{r} N^{-2/d} \right). \end{aligned}$$

In particular, the critical width is

$$s_{\text{crit}}(N, r) \asymp rN^{2/d-1/2}.$$

If $s \ll s_{\text{crit}}(N, r)$, the class is generically parametric; if $s \gg s_{\text{crit}}(N, r)$, it is generically nonparametric.

Proof. Combine Theorems 8.9 and 8.10. The critical width is obtained by balancing the two terms $sN^{-1/2}$ and $(s^2/r)N^{-2/d}$. \square

Remark 8.12. The mixed-scale theorem explains why thin translated annulus classes are nongeneric. Such classes can cancel the first-order term and exhibit a direct-versus-plug-in logarithmic gap far from the diagonal, but the full two-ball class does not: generically, the first-order mean term controls the narrow regime, while the second-order quadratic term controls the thick regime.

9 Exact thin-annulus classes away from the diagonal

The two-ball phase diagram describes the generic behavior of full product neighborhoods. There is, however, a smaller off-diagonal geometry in which the first-order term cancels exactly. On these translated thin annuli, direct estimation again beats plug-in by a logarithmic factor.

Throughout this section assume $d \geq 2$, $p \geq 1$, and $n = m = N$. Fix $0 < r \leq 1/10$.

Lemma 9.1 (Translated paired packing). *There exists $c_d > 0$ such that for every $M \geq 1$ and every $0 < \delta \leq c_d r M^{-1/(d-1)}$, one can choose points $z_1, \dots, z_M \in [0, 1]^d$ such that, with*

$$x_j^\pm := z_j \pm \frac{\delta}{2} e_2, \quad y_j^\pm := z_j + r e_1 \pm \frac{\delta}{2} e_2,$$

all four points lie in $[0, 1]^d$, and $\|z_i - z_j\|_2 \geq 8\delta$ whenever $i \neq j$.

Proof. Choose the anchors z_j on a grid of spacing 8δ inside a $(d-1)$ -dimensional box of side length comparable to r in the coordinates $2, \dots, d$. The bound $\delta \lesssim r M^{-1/(d-1)}$ guarantees room for M such points, and the translated pairs remain inside $[0, 1]^d$ when the constant is chosen small enough. \square

For $\theta, \eta \in [-1, 1]^M$, define

$$P_\theta := \frac{1}{2M} \sum_{j=1}^M ((1 + \theta_j) \delta_{x_j^+} + (1 - \theta_j) \delta_{x_j^-}),$$

$$Q_\eta := \frac{1}{2M} \sum_{j=1}^M ((1 + \eta_j) \delta_{y_j^+} + (1 - \eta_j) \delta_{y_j^-}),$$

and let q_θ, q_η be the corresponding laws on the alphabet $\mathcal{A}_M = \{(j, +), (j, -) : 1 \leq j \leq M\}$. Set

$$D_{r,\delta} := \sqrt{r^2 + \delta^2}, \quad a_{r,\delta,p} := D_{r,\delta}^p - r^p.$$

Proposition 9.2 (Exact translated paired reduction). *For every $\theta, \eta \in [-1, 1]^M$,*

$$W_p(P_\theta, Q_\eta)^p = r^p + a_{r,\delta,p} \text{TV}(q_\theta, q_\eta) = r^p + \frac{a_{r,\delta,p}}{2M} \sum_{j=1}^M |\theta_j - \eta_j|.$$

Proof. Within a fixed block j , same-sign moves have cost r^p , whereas cross-sign moves have cost $D_{r,\delta}^p$. If $i \neq j$, any source point in block i and target point in block j differ by $-r e_1$ plus an orthogonal component of norm at least 7δ , so their distance exceeds $D_{r,\delta}$. Because both marginals assign total mass $1/M$ to every block, the same cycle-elimination argument as in Theorem 3.2 shows that an optimal coupling is localized blockwise. Inside block j , exactly $|\theta_j - \eta_j|/(2M)$ units of mass must be matched cross-sign. Summing the resulting block costs proves the formula. \square

Lemma 9.3 (Quadratic annulus thickness). *If $0 < \delta \leq r$, then*

$$a_{r,\delta,p} \lesssim_p r^{p-2} \delta^2.$$

Moreover, for every $\theta, \eta \in [-1, 1]^M$,

$$r \leq W_p(P_\theta, Q_\eta) \leq r + C_p \frac{\delta^2}{r}.$$

Proof. Apply the mean value theorem to $t \mapsto (r^2 + t)^{p/2}$ on $[0, \delta^2]$. The bound for W_p then follows from $W_p(P_\theta, Q_\eta)^p \in [r^p, r^p + a_{r,\delta,p}]$ and another application of the mean value theorem to $x \mapsto x^{1/p}$. \square

Theorem 9.4 (Direct estimation on exact translated thin annuli). *Assume $\log N \leq C_0 \log(2M)$, where C_0 is the constant from [3], and let $0 < \delta \leq r$. Then there exists an estimator \widehat{W} such that*

$$\sup_{\theta, \eta \in [-1, 1]^M} \mathbb{E}(\widehat{W} - W_p(P_\theta, Q_\eta))^2 \lesssim_p r^{2-2p} a_{r, \delta, p}^2 \frac{M}{N \log N}.$$

In particular,

$$\sup_{\theta, \eta} \mathbb{E}(\widehat{W} - W_p(P_\theta, Q_\eta))^2 \lesssim_p \frac{\delta^4}{r^2} \cdot \frac{M}{N \log N}.$$

Proof. By Theorem 9.2, estimating $W_p(P_\theta, Q_\eta)^p$ is equivalent to estimating $\text{TV}(q_\theta, q_\eta)$ on an alphabet of size $2M$. Apply the Jiao–Han–Weissman estimator, clip it to $[0, 1]$, and then pass from $r^p + a_{r, \delta, p} \text{TV}$ to the distance itself via the map $x \mapsto x^{1/p}$, which is Lipschitz on $[r^p, r^p + a_{r, \delta, p}]$ with constant $\lesssim_p r^{1-p}$. The final display uses Theorem 9.3. \square

Proposition 9.5 (Plug-in remains logarithmically worse off the diagonal). *Assume $M \geq cN$ for a fixed constant $c > 0$, and let $P_0 := P_\theta$, $Q_0 := Q_\eta$ with $\theta = \eta = 0$. Then*

$$\mathbb{E}(W_p((P_0)_N, (Q_0)_N) - W_p(P_0, Q_0))^2 \gtrsim_{c, p} r^{2-2p} a_{r, \delta, p}^2 \asymp_{c, p} \frac{\delta^4}{r^2}$$

whenever $0 < \delta \leq r$.

Proof. The true distance equals $W_p(P_0, Q_0) = r$. After subtracting the baseline r^p from every source-target edge cost, same-label edges inside one block have modified cost 0, whereas every other edge has modified cost at least $a_{r, \delta, p}$. Hence

$$W_p((P_0)_N, (Q_0)_N)^p - r^p \geq a_{r, \delta, p} \text{TV}(\widehat{q}, \widehat{r}),$$

where \widehat{q} and \widehat{r} are the empirical histograms on \mathcal{A}_M . Since $\delta \leq r$, the map $t \mapsto (r^p + t)^{1/p} - r$ is bounded below on $[0, a_{r, \delta, p}]$ by a positive multiple of $r^{1-p}t$. Therefore

$$W_p((P_0)_N, (Q_0)_N) - r \gtrsim_p r^{1-p} a_{r, \delta, p} \text{TV}(\widehat{q}, \widehat{r}).$$

Taking expectations and applying Theorem 5.13 on the alphabet \mathcal{A}_M yields a positive lower bound on $\mathbb{E} \text{TV}(\widehat{q}, \widehat{r})$. Jensen’s inequality then gives the squared lower bound. \square

Corollary 9.6 (A geometric off-diagonal logarithmic gap). *Fix $0 < r \leq 1/10$, assume $d \geq 2$, $p \geq 1$, and let $n = m = N$. For all sufficiently large N , there exists an exact Euclidean class $\mathcal{A}_{N, r}^{\text{off}} \subset \mathcal{P}_d^2$ such that every $(P, Q) \in \mathcal{A}_{N, r}^{\text{off}}$ satisfies*

$$r \leq W_p(P, Q) \leq r + C_{d, p} r N^{-2/(d-1)},$$

while a direct estimator achieves

$$\sup_{(P, Q) \in \mathcal{A}_{N, r}^{\text{off}}} \mathbb{E}(\widehat{W} - W_p(P, Q))^2 \lesssim_{d, p} r^2 N^{-4/(d-1)} (\log N)^{-1/p},$$

and plug-in obeys the lower bound

$$\sup_{(P, Q) \in \mathcal{A}_{N, r}^{\text{off}}} \mathbb{E}(W_p(P_N, Q_N) - W_p(P, Q))^2 \gtrsim_{d, p} r^2 N^{-4/(d-1)}.$$

Proof. Take $M \asymp N$ and $\delta \asymp rN^{-1/(d-1)}$. Then Theorems 9.3 to 9.5 yield the stated upper and lower bounds, while Theorem 9.3 controls the location of the class in a thin annulus about radius r . \square

10 Known-reference lower bound and sample-size imbalance

The exact local theorem has a useful one-sided counterpart: the hard supercritical family already survives when one marginal is known exactly. This is the lower-bound ingredient needed to propagate any balanced supercritical upper bound to arbitrary sample sizes.

For $n \geq 1$, define the one-sided absolute and squared risks

$$\mathfrak{M}_{n,d,p}^{\text{abs},\rightarrow} := \inf_{\widehat{W}} \sup_{P,Q \in \mathcal{P}_d} \mathbb{E}_{P^{\otimes n}} |\widehat{W}(X_1, \dots, X_n; Q) - W_p(P, Q)|,$$

$$\mathfrak{M}_{n,d,p}^{\text{sq},\rightarrow} := \inf_{\widehat{W}} \sup_{P,Q \in \mathcal{P}_d} \mathbb{E}_{P^{\otimes n}} (\widehat{W}(X_1, \dots, X_n; Q) - W_p(P, Q))^2,$$

where the estimator may depend measurably on the exact reference measure Q .

Proposition 10.1 (Exact reference dominates sampled reference). *For every $n \geq 1$, $d \geq 1$, and $p \geq 1$,*

$$\mathfrak{M}_{n,d,p}^{\text{abs},\rightarrow} \leq \mathfrak{M}_{n,n,d,p}^{\text{abs}}, \quad \mathfrak{M}_{n,d,p}^{\text{sq},\rightarrow} \leq \mathfrak{M}_{n,n,d,p}^{\text{sq}}.$$

Proof. Take any two-sample estimator $\widehat{W}^{(2)}$ based on n samples from each marginal. Given an exact reference measure Q and an X -sample from P , simulate an independent auxiliary sample $Y_1, \dots, Y_n \sim Q^{\otimes n}$ and define

$$\widetilde{W}(X_1, \dots, X_n; Q) := \widehat{W}^{(2)}(X_1, \dots, X_n, Y_1, \dots, Y_n).$$

The joint law of $(X_1, \dots, X_n, Y_1, \dots, Y_n)$ is exactly $P^{\otimes n} \otimes Q^{\otimes n}$, so the risk of \widetilde{W} in the known-reference experiment equals the risk of $\widehat{W}^{(2)}$ in the balanced two-sample experiment. Conditioning on the observed X -sample and using Jensen's inequality removes the auxiliary randomization without increasing either absolute or squared loss. Taking infima proves the claim. \square

For $d > 2p$, set $\eta_n := (n \log n)^{-1/d}$. For a reference measure $Q \in \mathcal{P}_d$ and $A > 0$, define the one-sided local class

$$\mathcal{L}_{A,n}^{\rightarrow}(Q) := \{P \in \mathcal{P}_d : W_p(P, Q) \leq A\eta_n\}.$$

Theorem 10.2 (Known-reference local supercritical lower bound). *Assume $d > 2p$. There exist constants $A_1 = A_1(d, p)$, $c_1 = c_1(d, p) > 0$, and $n_1 = n_1(d, p)$ such that for all $n \geq n_1$ there exists a reference measure $Q_n \in \mathcal{P}_d$ with the following property: every estimator \widehat{W} based on n samples from P and exact knowledge of Q_n satisfies*

$$\sup_{P \in \mathcal{L}_{A_1,n}^{\rightarrow}(Q_n)} \mathbb{P}_{P^{\otimes n}} \left(|\widehat{W}(X_1, \dots, X_n; Q_n) - W_p(P, Q_n)| \geq c_1 \eta_n \right) \geq c_1.$$

Consequently,

$$\mathfrak{M}_{n,d,p}^{\text{abs},\rightarrow} \gtrsim_{d,p} \eta_n, \quad \mathfrak{M}_{n,d,p}^{\text{sq},\rightarrow} \gtrsim_{d,p} \eta_n^2.$$

Proof. Rerun the proof of Theorem 5.11 with $N = n$, but keep the reference marginal fixed. Using the same notation as there, let u be the uniform law on $[m]$, choose $m \asymp n \log n$, let $F : [m] \rightarrow [0, 1]^d$ be the random map from [7, Proposition 9], and set

$$\mu_{F,q} := F\#q, \quad \nu_F := F\#u.$$

The proof of [7, Theorem 11], together with the localization carried out in the proof of Theorem 5.11, yields constants A_1 , $\Delta_n \asymp \eta_n$, and a good event of probability at least 0.9 on which

$$\begin{aligned} W_p(\mu_{F,q}, \nu_F) &\leq \Delta_n & (q \in D_{m,\delta}^-), \\ 3\Delta_n &\leq W_p(\mu_{F,q}, \nu_F) \leq A_1\eta_n & (q \in D_m^+). \end{aligned}$$

In particular, all hard alternatives lie in the one-sided local class $\mathcal{L}_{A_1,n}^{\rightarrow}(\nu_F)$. Applying the same threshold test as in the proof of Theorem 5.11 now gives a constant probability lower bound for estimation with exact reference ν_F . Fixing one realization of F for which that bound holds and setting $Q_n := \nu_F$ proves the claim. The risk lower bounds follow from the probability bound exactly as in Theorem 5.11. \square

Remark 10.3. Theorem 10.2 is the part of the arbitrary-sample-size theory that is independent of the singular and boundary closures. Once a balanced supercritical upper bound of order $(N \log N)^{-1/d}$ is available, the general unbalanced upper bound follows by discarding surplus observations, while the lower bound follows by restricting to the smaller-sample side and fixing the other marginal to Q_N . Thus the exact reference experiment shows that sample-size imbalance cannot improve the minimax order beyond the smaller sample size $N = n \wedge m$.

11 Critical linearization, semiconcavity, and localization to nearest-neighbor offsets

The local and positive-distance theorems determine the diagonal regime and the generic smooth off-diagonal geometry. To complete the unrestricted supercritical theory it remains to control the residual off-diagonal layer on the critical grid. The decisive step is to linearize the Euclidean cost on dyadic macro-edges and then eliminate all second-order pieces that fall into one compact semiconcave class.

Throughout this section, for $p > 1$ set

$$\beta_p := p \wedge 2, \quad \eta_N := (N \log N)^{-1/d},$$

and define the dimension-only comparison scale

$$\rho_d(N) := \begin{cases} N^{-1/2}, & d \leq 3, \\ N^{-1/2} \log N, & d = 4, \\ N^{-2/d}, & d \geq 5. \end{cases}$$

Let $Q_{\square} := [-1, 1]^d$ and, for a bounded cost g on $Q_{\square} \times Q_{\square}$, write

$$T_g(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int g(u, v) d\pi(u, v).$$

Proposition 11.1 (Semiconcave branching principle). *Let \mathcal{G} be a compact class of bounded nonnegative transport costs on $Q_\square \times Q_\square$ that are uniformly Lipschitz and uniformly semiconcave in each variable separately. Then*

$$\sup_{g \in \mathcal{G}} \sup_{\mu, \nu \in \mathcal{P}(Q_\square)} \mathbb{E} |T_g(\widehat{\mu}_N, \widehat{\nu}_N) - T_g(\mu, \nu)| \lesssim_d \rho_d(N),$$

and

$$\sup_{g \in \mathcal{G}} \sup_{\mu, \nu \in \mathcal{P}(Q_\square)} \mathbb{E} (T_g(\widehat{\mu}_N, \widehat{\nu}_N) - T_g(\mu, \nu))^2 \lesssim_d \rho_d(N)^2 + N^{-1}.$$

In the supercritical regime $d > 2p$, one has $\rho_d(N) = o(\eta_N)$ and $\rho_d(N)^2 + N^{-1} = o(\eta_N^2)$.

Proof. After normalization, every dual-optimal potential for a cost in \mathcal{G} belongs on each marginal side to one fixed uniformly bounded, uniformly Lipschitz, uniformly semiconcave class on the cube Q_\square . Subtracting the corresponding quadratic semiconcavity modulus converts that class into a uniformly bounded Lipschitz concave family. Bronshtein's entropy theorem for convex/concave functions on a fixed d -dimensional body, together with the L_1 -covering summary in Guntuboyina and Sen [4, 5], therefore yields the metric-entropy exponent

$$\log \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \lesssim_d \varepsilon^{-d/2}$$

for the normalized dual envelope \mathcal{F} . That entropy exponent gives the deviation scale $\rho_d(N)$ through the standard entropy-integral bound for empirical processes, while bounded differences provide the corresponding second-moment estimate $\rho_d(N)^2 + N^{-1}$. The final little- o comparison follows immediately from $d > 2p$. \square

Proposition 11.2 (Affine first-order transport costs are parametric). *Let $\mathcal{A} \subset \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ be compact, and for $\alpha = (a, b, c_0) \in \mathcal{A}$ define*

$$g_\alpha(u, v) := a \cdot u + b \cdot v + c_0, \quad (u, v) \in Q_\square \times Q_\square.$$

Then

$$\sup_{\alpha \in \mathcal{A}} \sup_{\mu, \nu \in \mathcal{P}(Q_\square)} \mathbb{E} |T_{g_\alpha}(\widehat{\mu}_N, \widehat{\nu}_N) - T_{g_\alpha}(\mu, \nu)| \lesssim_{\mathcal{A}, d} N^{-1/2},$$

and

$$\sup_{\alpha \in \mathcal{A}} \sup_{\mu, \nu \in \mathcal{P}(Q_\square)} \mathbb{E} (T_{g_\alpha}(\widehat{\mu}_N, \widehat{\nu}_N) - T_{g_\alpha}(\mu, \nu))^2 \lesssim_{\mathcal{A}, d} N^{-1}.$$

In particular, on the critical grid the first-order affine part of the linearized cost contributes at most $C_{d,p} N^{-1/2}$ in absolute loss and $C_{d,p} N^{-1}$ in squared loss, uniformly over all finite and compact affine branching families generated by the Euclidean linearization.

Proof. For every affine cost g_α , the transport value is independent of the coupling:

$$T_{g_\alpha}(\mu, \nu) = a \cdot \int u d\mu(u) + b \cdot \int v d\nu(v) + c_0.$$

Write

$$m_\mu := \int u d\mu(u), \quad m_\nu := \int v d\nu(v), \quad \widehat{m}_\mu := \int u d\widehat{\mu}_N(u), \quad \widehat{m}_\nu := \int v d\widehat{\nu}_N(v),$$

and let

$$A := \sup_{\alpha=(a,b,c_0) \in \mathcal{A}} (\|a\|_2 + \|b\|_2) < \infty.$$

Then

$$|T_{g_\alpha}(\widehat{\mu}_N, \widehat{\nu}_N) - T_{g_\alpha}(\mu, \nu)| \leq A\|\widehat{m}_\mu - m_\mu\|_2 + A\|\widehat{m}_\nu - m_\nu\|_2.$$

Since $Q_\square = [-1, 1]^d$, every coordinate of $U \sim \mu$ and $V \sim \nu$ is bounded by 1 in absolute value. Therefore

$$\mathbb{E}\|\widehat{m}_\mu - m_\mu\|_2^2 = \sum_{j=1}^d \text{Var}\left(\frac{1}{N} \sum_{i=1}^N U_{i,j}\right) \leq \frac{d}{N}, \quad \mathbb{E}\|\widehat{m}_\nu - m_\nu\|_2^2 \leq \frac{d}{N}.$$

By Cauchy–Schwarz,

$$\sup_{\alpha, \mu, \nu} \mathbb{E}|T_{g_\alpha}(\widehat{\mu}_N, \widehat{\nu}_N) - T_{g_\alpha}(\mu, \nu)| \leq 2A\sqrt{\frac{d}{N}} \lesssim_{\mathcal{A}, d} N^{-1/2},$$

and similarly

$$\mathbb{E}(T_{g_\alpha}(\widehat{\mu}_N, \widehat{\nu}_N) - T_{g_\alpha}(\mu, \nu))^2 \leq 2A^2\left(\mathbb{E}\|\widehat{m}_\mu - m_\mu\|_2^2 + \mathbb{E}\|\widehat{m}_\nu - m_\nu\|_2^2\right) \lesssim_{\mathcal{A}, d} N^{-1}.$$

For the last sentence, observe that the normalized affine coefficients produced by the critical Euclidean linearization lie in one compact coefficient set depending only on d and p . \square

For $p > 1$ and dyadic level $J \geq 1$, let c_R denote the center of a level- J dyadic cube R . Define the piecewise affine linearized cost by

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

for $x \in R$ and $y \in S$, and set

$$T_{J,p}^{\text{lin}}(P, Q) := \inf_{\pi \in \Pi(P, Q)} \int c_{J,p}^{\text{lin}}(x, y) d\pi(x, y).$$

Proposition 11.3 (Critical linearization reduction for $p > 1$). *Assume $d > 2p$, $p > 1$, and $n = m = N$. Choose $J = J_{\text{lin}}(N)$ so that*

$$2^{-J\beta_p} \asymp \eta_N^p.$$

Then

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

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

Proof. On each dyadic macro-edge, Taylor expansion of $h_p(z) = \|z\|_2^p$ around the macro-center gives a pointwise error of order $2^{-J\beta_p}$, uniformly over all local positions inside the two cubes. Integrating this bound against an arbitrary coupling and taking the infimum proves the transport-cost statement. When $1 < p \leq 2$, one has $\beta_p = p$, so $2^{-Jp} \asymp \eta_N^p$ is equivalent to $2^{Jd} \asymp N \log N$. \square

Proposition 11.4 (Support-sparse and dyadically thin families are subcritical). *Assume $d > 2p$.*

(i) If $|\text{supp}(P) \cup \text{supp}(Q)| \leq S_N$, then

$$\sup_{\substack{P, Q \in \mathcal{P}_d: \\ |\text{supp}(P) \cup \text{supp}(Q)| \leq S_N}} \mathbb{E} |W_p(P_N, Q_N) - W_p(P, Q)| \lesssim_{d,p} \left(\frac{S_N}{N}\right)^{1/(2p)}.$$

In particular, every critical-grid family with

$$S_N \lesssim N^{1-2p/d} (\log N)^{-2p/d}$$

is already below the target scale η_N .

(ii) If the union support lies in a set $A \subset [0, 1]^d$ whose dyadic profile satisfies

$$\#\{R \in \mathcal{D}_j : R \cap A \neq \emptyset\} \lesssim 2^{js} \quad (j \geq 0)$$

for some $s < d$, then the empirical plug-in estimator achieves the intrinsic rate $N^{-1/(2p)}$, $N^{-1/(2p)}(\log N)^{1/p}$, or $N^{-1/s}$ according as $s < 2p$, $s = 2p$, or $s > 2p$. In particular, every critical-grid family of macro-dimension strictly smaller than d is already subcritical.

Proof. The finite-state estimate in (i) follows from the Lipschitz control of the Wasserstein functional on a support of size S_N and the L^2 fluctuations of empirical histograms. The dyadic bound in (ii) follows by summing empirical mass imbalances over the occupied dyadic cubes and exploiting the profile exponent s . \square

Theorem 11.5 (Finite-offset localization of the unresolved supercritical layer). *Assume $d > 2p$, $p > 1$, and $n = m = N$. After the critical linearization reduction of Theorem 11.3, the parametric first-order affine elimination of Theorem 11.2, the diagonal theorem Theorem 5.12, the semiconcave branching principle of Theorem 11.1, and the support-complexity eliminations of Theorem 11.4, any remaining unresolved second-order contribution is generated by the finite nearest-neighbor offset set*

$$\mathcal{N}_d := \{-1, 0, 1\}^d \setminus \{0\}.$$

More precisely:

- (i) same-cell transport is diagonal at scale 2^{-J} and is therefore absorbed by Theorem 5.12;
- (ii) every far offset m with $\|m\|_\infty \geq 2$ is semiconcave after normalization and hence contributes at most $O(\rho_d(N))$ in absolute loss and $O(\rho_d(N)^2 + N^{-1})$ in squared loss;
- (iii) any hard family must have full macro-support and must live on the finite nearest-neighbor macrograph determined by \mathcal{N}_d .

Proof. Write $x = c_R + h_J u$ and $y = c_S + h_J v$ with $u, v \in Q_\square$, and set $m := h_J^{-1}(c_R - c_S) \in \mathbb{Z}^d$. Then the exact Taylor decomposition of $\|x - y\|_2^p$ on the macro-edge (R, S) has the form

$$\|x - y\|_2^p = h_J^p \|m\|_2^p + h_J^p \nabla h_p(m) \cdot (u - v) + h_J^p \|m\|_2^{p-2} \Psi_m(u - v),$$

where the first term is deterministic, the second is affine, and the third is a normalized second-order residual. The affine term is already parametric by Theorem 11.2 and is therefore subcritical. If $m = 0$, then $R = S$ and $\|x - y\|_2 \leq \sqrt{d} h_J$, so the contribution is diagonal. If $\|m\|_\infty \geq 2$, the normalized residual family lies in one compact semiconcave class and

is subcritical by Theorem 11.1. Finally, Theorem 11.4 removes all support-sparse and dyadically thin macro-geometries. The only unresolved offsets are therefore the finitely many nearest-neighbor ones in \mathcal{N}_d . \square

12 Balanced supercritical closure for $p > 1$

The decisive point in the closing argument is that the terminal reduced models must first be controlled at the level of the transport cost $W_p(P, Q)^p$; only after that does one pass to the distance $W_p(P, Q)$ through the $1/p$ -Hölder map $t \mapsto t^{1/p}$. This bookkeeping is straightforward in the semiconcave range $p \geq 2$, and it is essential in the singular range $1 < p < 2$, where the exact dyadic prefactors carried by the terminal blocks must be restored before the global estimate becomes visible.

12.1 The semiconcave range $p \geq 2$

Proposition 12.1 (Nearest-neighbor residuals form one compact semiconcave class for $p \geq 2$). *Assume $p \geq 2$. For every nearest-neighbor offset $m \in \mathcal{N}_d$, define the normalized residual*

$$\Psi_m(z) := \|m\|_2^{2-p} \left(\|m+z\|_2^p - \|m\|_2^p - \nabla h_p(m) \cdot z \right), \quad z \in Q_\square,$$

where $h_p(z) := \|z\|_2^p$. Then the family $\{\Psi_m : m \in \mathcal{N}_d\}$ is contained in one fixed compact separately semiconcave class on Q_\square .

Proof. For $p \geq 2$, the map $h_p(z) = \|z\|_2^p$ belongs to $C^2(\mathbb{R}^d)$. For every $m \in \mathcal{N}_d$, the translated cube $m + Q_\square$ lies inside the fixed compact set $[-2, 2]^d$, so the Hessian of h_p is uniformly bounded there. Taylor's integral remainder formula therefore gives uniform boundedness, uniform Lipschitz control, and uniform semiconcavity for Ψ_m on Q_\square , with constants depending only on d and p . Since \mathcal{N}_d is finite, compactness is immediate. \square

Proposition 12.2 (Cost-level closure under semiconcave nearest-neighbor control). *Assume $p \geq 2$, $d > 2p$, and $n = m = N$. Suppose that every nearest-neighbor residual kernel belongs to one fixed compact separately semiconcave class on Q_\square . Then there exists an estimator \widehat{T}_N of $W_p(P, Q)^p$ such that*

$$\sup_{P, Q \in \mathcal{P}_d} \mathbb{E} |\widehat{T}_N - W_p(P, Q)^p| \lesssim_{d,p} \eta_N^p, \quad \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} (\widehat{T}_N - W_p(P, Q)^p)^2 \lesssim_{d,p} \eta_N^{2p}.$$

Consequently, after clipping \widehat{T}_N to $[0, d^{p/2}]$ and setting $\widehat{W}_N := \widehat{T}_N^{1/p}$, one has

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

Proof. Let $J = J_{\text{lin}}(N)$ be the critical level from Theorem 11.3, and write $h_J := 2^{-J}$. By Theorem 11.3, it is enough to estimate the linearized transport cost $T_{J,p}^{\text{lin}}(P, Q)$ at absolute scale η_N^p and squared scale η_N^{2p} .

For a nonzero macro-edge offset m , the exact Taylor decomposition used in the proof of Theorem 11.5 reads

$$\|x - y\|_2^p = h_J^p \|m\|_2^p + h_J^p \nabla h_p(m) \cdot (u - v) + h_J^p \|m\|_2^{p-2} \Psi_m(u - v),$$

where $x = c_R + h_J u$, $y = c_S + h_J v$, and $u, v \in Q_\square$.

We inspect the resulting pieces one by one.

Same-cell transport. If $m = 0$, then $R = S$ and $\|x - y\|_2 \leq \sqrt{d} h_J$, so the same-cell contribution is $O(h_J^p)$ in absolute loss and $O(h_J^{2p})$ in squared loss. When $p \geq 2$, $\beta_p = 2$, hence $h_J^2 \asymp \eta_N^p$, and therefore $h_J^p \leq h_J^2 \asymp \eta_N^p$ and $h_J^{2p} \leq h_J^4 \asymp \eta_N^{2p}$.

First-order affine part. For every level- J offset m , one has

$$h_J^p \|\nabla h_p(m)\|_2 \lesssim_{d,p} h_J,$$

because $\|\nabla h_p(m)\|_2 \asymp_{d,p} \|m\|_2^{p-1}$ and $\|m\|_2 \lesssim_d h_J^{-1}$ on the level- J grid. After dividing by this deterministic prefactor, the remaining local cost is affine on $Q_\square \times Q_\square$, so Theorem 11.2 gives normalized empirical fluctuation $O(N^{-1/2})$ in absolute loss and $O(N^{-1})$ in squared loss. Hence the lifted affine contribution is

$$O(h_J N^{-1/2}) = o(\eta_N^p)$$

in absolute loss and

$$O(h_J^2 N^{-1}) = o(\eta_N^{2p})$$

in squared loss. Indeed $h_J \asymp \eta_N^{p/2}$, so the two ratios are $N^{-1/2} \eta_N^{-p/2} \rightarrow 0$ and $N^{-1} \eta_N^{-p} \rightarrow 0$ because $d > 2p$.

Far offsets. If $\|m\|_\infty \geq 2$, then Theorem 11.5 places the normalized residuals in a compact semiconcave class. Their normalized empirical deviation is therefore $O(\rho_d(N))$ in absolute loss and $O(\rho_d(N)^2 + N^{-1})$ in squared loss by Theorem 11.1. Since

$$h_J^p \|m\|_2^{p-2} \lesssim_{d,p} h_J^2 \asymp \eta_N^p \quad (p \geq 2),$$

the lifted far-offset contribution is again $o(\eta_N^p)$ and $o(\eta_N^{2p})$.

Nearest-neighbor offsets. Under the present hypothesis, the normalized nearest-neighbor residuals satisfy the same semiconcavity conclusion. Because $\|m\|_2 \asymp_d 1$ for $m \in \mathcal{N}_d$, the prefactor $h_J^p \|m\|_2^{p-2}$ is $O(h_J^p) \leq O(h_J^2) \asymp O(\eta_N^p)$. Applying Theorem 11.1 on the normalized local cube therefore yields a lifted nearest-neighbor contribution $o(\eta_N^p)$ in absolute loss and $o(\eta_N^{2p})$ in squared loss.

Collecting the same-cell term, the lifted affine term, the lifted far-offset term, and the lifted nearest-neighbor term yields an estimator \widehat{T}_N with

$$\sup_{P, Q \in \mathcal{P}_d} \mathbb{E} |\widehat{T}_N - W_p(P, Q)^p| \lesssim_{d,p} \eta_N^p, \quad \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} (\widehat{T}_N - W_p(P, Q)^p)^2 \lesssim_{d,p} \eta_N^{2p}.$$

Finally, clip \widehat{T}_N to $[0, d^{p/2}]$ and define $\widehat{W}_N := \widehat{T}_N^{1/p}$. Since $t \mapsto t^{1/p}$ is $1/p$ -Hölder on $[0, \infty)$,

$$|\widehat{W}_N - W_p(P, Q)| \leq |\widehat{T}_N - W_p(P, Q)^p|^{1/p}.$$

Lyapunov's inequality therefore gives the distance-level bounds

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

□

Theorem 12.3 (Exact unrestricted balanced supercritical minimax law for $p \geq 2$). *Assume $p \geq 2$, $d > 2p$, and $n = m = N$. Then*

$$\inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} |\widehat{W} - W_p(P, Q)| \asymp_{d,p} (N \log N)^{-1/d},$$

and

$$\inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} (\widehat{W} - W_p(P, Q))^2 \asymp_{d,p} (N \log N)^{-2/d}.$$

Proof. The lower bounds are exactly Theorem 2.1(iv). By Theorem 12.1, the nearest-neighbor family satisfies the hypothesis of Theorem 12.2. That proposition therefore yields matching upper bounds. \square

12.2 The singular range $1 < p < 2$

Proposition 12.4 (Corner normal form). *Assume $1 < p < 2$. Up to signed permutations and affine perturbations, every nearest-neighbor residual kernel reduces to one of the d canonical corner kernels*

$$w \longmapsto k^{(2-p)/2} \|w\|_2^p, \quad w \in [0, 2]^k \times [-1, 1]^{d-k}, \quad k = 1, \dots, d.$$

Proof. Every nearest-neighbor vector can be sent by a signed permutation to $m^{(k)} := e_1 + \dots + e_k$ for some $k \in \{1, \dots, d\}$. Orthogonal invariance of the Euclidean norm implies that the normalized residual transforms equivariantly under this symmetry. Writing $z = -m^{(k)} + w$ with $w \in [0, 2]^k \times [-1, 1]^{d-k}$ and substituting into the exact Taylor remainder then yields

$$\Psi_{m^{(k)}}(-m^{(k)} + w) = k^{(2-p)/2} \|w\|_2^p - p \sum_{j=1}^k w_j + (p-1)k.$$

Removing the affine term and the additive constant leaves the stated radial corner kernel. \square

Proposition 12.5 (Reverse bias for fixed bounded continuous costs). *Let X and Y be compact metric spaces, let $c : X \times Y \rightarrow \mathbb{R}$ be bounded and continuous, and let $\widehat{\mu}_N, \widehat{\nu}_N$ be the empirical measures generated by $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Then*

$$T_c(\mu, \nu) \leq \mathbb{E} T_c(\widehat{\mu}_N, \widehat{\nu}_N).$$

Proof. By Kantorovich duality on compact metric spaces, there exist bounded Borel potentials $\varphi : X \rightarrow \mathbb{R}$ and $\psi : Y \rightarrow \mathbb{R}$ such that

$$\varphi(x) + \psi(y) \leq c(x, y) \quad (x \in X, y \in Y)$$

and

$$T_c(\mu, \nu) = \int_X \varphi d\mu + \int_Y \psi d\nu.$$

The same admissible pair (φ, ψ) is feasible for every empirical pair $(\widehat{\mu}_N, \widehat{\nu}_N)$, so

$$T_c(\widehat{\mu}_N, \widehat{\nu}_N) \geq \int_X \varphi d\widehat{\mu}_N + \int_Y \psi d\widehat{\nu}_N.$$

Taking expectations and using $\mathbb{E} \widehat{\mu}_N = \mu$ and $\mathbb{E} \widehat{\nu}_N = \nu$ gives the claim. \square

Lemma 12.6 (Finite-dimensional switch parameters contribute only subcritical fluctuations). *Let X and Y be compact metric spaces, let $f_1, \dots, f_L : X \times Y \rightarrow \mathbb{R}$ be bounded continuous, let $\Lambda = [-M, M]^L$, and define*

$$g_\lambda := \min_{1 \leq j \leq L} (f_j - \lambda_j), \quad Z_{N,\lambda} := T_{g_\lambda}(\widehat{\mu}_N, \widehat{\nu}_N) - \mathbb{E}T_{g_\lambda}(\widehat{\mu}_N, \widehat{\nu}_N).$$

Set

$$B := \max_{1 \leq j \leq L} \|f_j\|_{L^\infty(X \times Y)} + M.$$

Then for every $\varepsilon \in (0, 1]$,

$$\sup_{\mu, \nu} \mathbb{E} \sup_{\lambda \in \Lambda} |Z_{N,\lambda}| \lesssim_{B,L,M} \varepsilon + N^{-1/2} \sqrt{\log(e/\varepsilon)},$$

and

$$\sup_{\mu, \nu} \mathbb{E} \sup_{\lambda \in \Lambda} |Z_{N,\lambda}|^2 \lesssim_{B,L,M} \varepsilon^2 + N^{-1} \log(e/\varepsilon).$$

Proof. For $\lambda, \lambda' \in \Lambda$,

$$\|g_\lambda - g_{\lambda'}\|_{L^\infty(X \times Y)} \leq \|\lambda - \lambda'\|_\infty.$$

Hence for every pair of probability measures $\alpha \in \mathcal{P}(X)$, $\beta \in \mathcal{P}(Y)$,

$$|T_{g_\lambda}(\alpha, \beta) - T_{g_{\lambda'}}(\alpha, \beta)| \leq \|\lambda - \lambda'\|_\infty.$$

Choose an ε -net $\Lambda_\varepsilon \subset \Lambda$ in the sup norm with $|\Lambda_\varepsilon| \leq C_{L,M} \varepsilon^{-L}$. Then every $\lambda \in \Lambda$ admits $\lambda' \in \Lambda_\varepsilon$ such that

$$|Z_{N,\lambda}| \leq |Z_{N,\lambda'}| + 2\varepsilon,$$

so

$$\sup_{\lambda \in \Lambda} |Z_{N,\lambda}| \leq \max_{\lambda' \in \Lambda_\varepsilon} |Z_{N,\lambda'}| + 2\varepsilon.$$

Fix $\lambda' \in \Lambda_\varepsilon$. Because $\|g_{\lambda'}\|_{L^\infty} \leq B$, replacing one source sample or one target sample changes $T_{g_{\lambda'}}(\widehat{\mu}_N, \widehat{\nu}_N)$ by at most $2B/N$. McDiarmid's inequality therefore yields

$$\mathbb{P}(|Z_{N,\lambda'}| \geq t) \leq 2 \exp\left(-\frac{Nt^2}{4B^2}\right) \quad (t > 0).$$

Applying the union bound over Λ_ε and integrating the tail gives

$$\mathbb{E} \max_{\lambda' \in \Lambda_\varepsilon} |Z_{N,\lambda'}| \lesssim_{B,L,M} N^{-1/2} \sqrt{\log(e/\varepsilon)}$$

and

$$\mathbb{E} \max_{\lambda' \in \Lambda_\varepsilon} |Z_{N,\lambda'}|^2 \lesssim_{B,L,M} N^{-1} \log(e/\varepsilon).$$

Combining these estimates with the approximation error 2ε proves the lemma. \square

Theorem 12.7 (Finite singular switch envelopes close at the branchwise rate). *Assume $1 < p < 2$ and $d > 2p$. Let $K_X, K_Y \subset \mathbb{R}^d$ be fixed compact convex sets, let f_1, \dots, f_L be any finite family of translated corner kernels arising from Theorem 12.4 on $K_X \times K_Y$, let $\Lambda = [-M, M]^L$, and define*

$$g_\lambda := \min_{1 \leq j \leq L} (f_j - \lambda_j), \quad T_\lambda := T_{g_\lambda}, \quad \widehat{T}_{\lambda,N} := T_{g_\lambda}(\widehat{\mu}_N, \widehat{\nu}_N).$$

Then

$$\sup_{\lambda \in \Lambda} \sup_{\mu \in \mathcal{P}(K_X), \nu \in \mathcal{P}(K_Y)} \mathbb{E} |\widehat{T}_{\lambda, N} - T_{\lambda}(\mu, \nu)| \lesssim_{d, p, K_X, K_Y, L, M} N^{-p/d},$$

and

$$\sup_{\lambda \in \Lambda} \sup_{\mu \in \mathcal{P}(K_X), \nu \in \mathcal{P}(K_Y)} \mathbb{E} (\widehat{T}_{\lambda, N} - T_{\lambda}(\mu, \nu))^2 \lesssim_{d, p, K_X, K_Y, L, M} N^{-2p/d}.$$

Proof. Write $a := p/d \in (0, 1/2)$. Each translated corner kernel is an individual singular p -kernel, so on the fixed compact convex class $K_X \times K_Y$ its empirical transport cost already satisfies absolute risk $O(m^{-a})$ and squared risk $O(m^{-2a})$.

Fix $\lambda \in \Lambda$ and $\mu \in \mathcal{P}(K_X)$, $\nu \in \mathcal{P}(K_Y)$. Choose an optimal coupling π for the envelope cost g_{λ} , and fix a measurable tie-breaking of the active branches $\{j : g_{\lambda}(u, v) = f_j(u, v) - \lambda_j\}$. This decomposes π into branch measures $\pi = \sum_{j=1}^L \pi_j$ with masses $t_j := \pi_j(K_X \times K_Y)$ and marginals $\mu_j := (\text{pr}_1)_{\#} \pi_j$, $\nu_j := (\text{pr}_2)_{\#} \pi_j$. Writing $\bar{\mu}_j := t_j^{-1} \mu_j$ and $\bar{\nu}_j := t_j^{-1} \nu_j$ on the event $t_j > 0$, one has

$$T_{\lambda}(\mu, \nu) = \sum_{j=1}^L \left(\int f_j d\pi_j - \lambda_j t_j \right) \geq \sum_{j=1}^L t_j T_{f_j}(\bar{\mu}_j, \bar{\nu}_j) - \lambda_j t_j.$$

Label the empirical X - and Y -samples independently according to the branch weights induced by (μ_j) and (ν_j) , let N_j and M_j be the corresponding counts, and thin each branch to equal size $m_j := \min\{N_j, M_j\}$. This produces branchwise empirical measures

$$\tilde{\mu}_j = \frac{m_j}{N} \widehat{\mu}_{j, m_j}, \quad \tilde{\nu}_j = \frac{m_j}{N} \widehat{\nu}_{j, m_j},$$

together with leftover source and target masses of common total mass

$$R := 1 - \sum_{j=1}^L \frac{m_j}{N}.$$

Since $m_j/N = \min\{N_j/N, M_j/N\}$, both $|m_j/N - t_j|$ and R are bounded by finite sums of the multinomial deviations $|N_j/N - t_j|$ and $|M_j/N - t_j|$. Because $g_{\lambda} \leq f_j - \lambda_j$ for every j , the empirical envelope cost is bounded by the feasible empirical branching plan

$$\widehat{T}_{\lambda, N} \leq \sum_{j=1}^L \left(\frac{m_j}{N} T_{f_j}(\widehat{\mu}_{j, m_j}, \widehat{\nu}_{j, m_j}) - \lambda_j \frac{m_j}{N} \right) + C_{f, \Lambda} R,$$

where $C_{f, \Lambda}$ depends only on the sup norms of the f_j and on the parameter box Λ .

Subtracting the preceding lower bound for $T_{\lambda}(\mu, \nu)$, taking positive parts, and using that L is fixed yield

$$\begin{aligned} (\widehat{T}_{\lambda, N} - T_{\lambda}(\mu, \nu))_+ &\lesssim_{f, \Lambda, L} \sum_{j=1}^L \frac{m_j}{N} |T_{f_j}(\widehat{\mu}_{j, m_j}, \widehat{\nu}_{j, m_j}) - T_{f_j}(\bar{\mu}_j, \bar{\nu}_j)| \\ &\quad + \sum_{j=1}^L \left| \frac{m_j}{N} - t_j \right| + R. \end{aligned}$$

Since $(m_j/N)m_j^{-a} \leq N^{-a}$ and $(m_j/N)^2m_j^{-2a} \leq N^{-2a}$, the branchwise absolute and squared laws give

$$\sup_{\lambda, \mu, \nu} \mathbb{E}(\widehat{T}_{\lambda, N} - T_{\lambda}(\mu, \nu))_+ \lesssim_{d, p, K_X, K_Y, L, M} N^{-a} + N^{-1/2},$$

and

$$\sup_{\lambda, \mu, \nu} \mathbb{E}(\widehat{T}_{\lambda, N} - T_{\lambda}(\mu, \nu))_+^2 \lesssim_{d, p, K_X, K_Y, L, M} N^{-2a} + N^{-1},$$

because $\mathbb{E}|m_j/N - t_j| + \mathbb{E}R \lesssim_L N^{-1/2}$ and $\mathbb{E}|m_j/N - t_j|^2 + \mathbb{E}R^2 \lesssim_L N^{-1}$.

On the reverse-bias side, Theorem 12.5 implies

$$T_{\lambda}(\mu, \nu) \leq \mathbb{E}\widehat{T}_{\lambda, N} \quad \text{for every fixed } \lambda.$$

Thus only centered fluctuations remain. Since $\lambda \mapsto g_{\lambda}$ is sup-norm Lipschitz, Theorem 12.6 with $\varepsilon := N^{-a}$ yields

$$\sup_{\mu, \nu} \mathbb{E} \sup_{\lambda \in \Lambda} |\widehat{T}_{\lambda, N} - \mathbb{E}\widehat{T}_{\lambda, N}| \lesssim_{d, p, K_X, K_Y, L, M} N^{-1/2} \sqrt{\log N} + N^{-a} = O(N^{-a}),$$

and

$$\sup_{\mu, \nu} \mathbb{E} \sup_{\lambda \in \Lambda} |\widehat{T}_{\lambda, N} - \mathbb{E}\widehat{T}_{\lambda, N}|^2 \lesssim_{d, p, K_X, K_Y, L, M} N^{-1} \log N + N^{-2a} = O(N^{-2a}),$$

because $a < 1/2$.

Combining the one-sided upper-bias bounds, the automatic reverse bias, and the centered fluctuation bounds proves the theorem. \square

Proposition 12.8 (Cost-level closure of the lifted singular terminal layer). *Assume $1 < p < 2$, $d > 2p$, and $n = m = N$. Then there exists an estimator \widehat{T}_N of $W_p(P, Q)^p$ such that*

$$\sup_{P, Q \in \mathcal{P}_d} \mathbb{E}|\widehat{T}_N - W_p(P, Q)^p| \lesssim_{d, p} \eta_N^p, \quad \sup_{P, Q \in \mathcal{P}_d} \mathbb{E}(\widehat{T}_N - W_p(P, Q)^p)^2 \lesssim_{d, p} \eta_N^{2p}.$$

Consequently, after clipping \widehat{T}_N to $[0, d^{p/2}]$ and setting $\widehat{W}_N := \widehat{T}_N^{1/p}$, one has

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

Proof. Let $J = J_{\text{lin}}(N)$ be the critical level from Theorem 11.3, and write $h_J := 2^{-J}$. Since $1 < p < 2$, one has $\beta_p = p$, hence $h_J^p \asymp \eta_N^p$. By Theorem 11.3, it is enough to estimate the linearized transport cost $T_{J, p}^{\text{lin}}(P, Q)$ at absolute scale η_N^p and squared scale η_N^{2p} .

As in the proof of Theorem 12.2, each nonzero offset block has an exact decomposition into a deterministic baseline, an affine first-order term with prefactor $h_J^p \nabla h_p(m)$, and a normalized second-order residual with prefactor $h_J^p \|m\|_2^{p-2}$.

Same-cell transport. If $m = 0$, then $R = S$ and $\|x - y\|_2 \leq \sqrt{d} h_J$, so the same-cell contribution is $O(h_J^p) = O(\eta_N^p)$ in absolute loss and $O(h_J^{2p}) = O(\eta_N^{2p})$ in squared loss.

First-order affine part. Uniformly over level- J offsets,

$$h_J^p \|\nabla h_p(m)\|_2 \lesssim_{d, p} h_J.$$

After dividing by this deterministic prefactor, the remaining local cost is affine on $Q_\square \times Q_\square$, so Theorem 11.2 gives normalized empirical fluctuation $O(N^{-1/2})$ in absolute loss and $O(N^{-1})$ in squared loss. Hence the lifted affine contribution is

$$O(h_J N^{-1/2}) = o(\eta_N^p)$$

in absolute loss and

$$O(h_J^2 N^{-1}) = o(\eta_N^{2p})$$

in squared loss. Because $h_J \asymp \eta_N$ in the range $1 < p < 2$, the two ratios reduce to $N^{-1/2} \eta_N^{1-p} \rightarrow 0$ and $N^{-1} \eta_N^{2-2p} \rightarrow 0$, which follows from $d > 2p > 2(p-1)$.

Far offsets. If $\|m\|_\infty \geq 2$, Theorem 11.5 and Theorem 11.1 give normalized empirical deviation $O(\rho_d(N))$ in absolute loss and $O(\rho_d(N)^2 + N^{-1})$ in squared loss. Because $p-2 < 0$ and $\|m\|_2 \geq 1$,

$$h_J^p \|m\|_2^{p-2} \lesssim_{d,p} h_J^p \asymp \eta_N^p,$$

so the lifted far-offset contribution is $o(\eta_N^p)$ and $o(\eta_N^{2p})$.

Nearest-neighbor terminal layer. By Theorem 11.5 and Theorem 11.4, every unresolved singular contribution reduces to a finite corner-switch family supported on the nearest-neighbor macrograph. Its normalized empirical fluctuation is controlled by Theorem 12.7: $N^{-p/d}$ in absolute loss and $N^{-2p/d}$ in squared loss. For nearest-neighbor offsets, $\|m\|_2 \asymp_d 1$, so the exact dyadic prefactor satisfies

$$h_J^p \|m\|_2^{p-2} \asymp_{d,p} \eta_N^p.$$

Therefore the lifted terminal singular contribution is

$$O(\eta_N^p N^{-p/d}) = o(\eta_N^p)$$

in absolute loss and

$$O(\eta_N^{2p} N^{-2p/d}) = o(\eta_N^{2p})$$

in squared loss.

Combining the same-cell term, the lifted affine term, the lifted far-offset term, and the lifted terminal singular term yields an estimator \widehat{T}_N with the displayed cost-level bounds. The distance-level bounds follow from the same clipping and $1/p$ -Hölder passage as in the proof of Theorem 12.2. \square

Theorem 12.9 (Exact unrestricted balanced supercritical minimax law in the singular range). *Assume $1 < p < 2$, $d > 2p$, and $n = m = N$. Then*

$$\inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} |\widehat{W} - W_p(P, Q)| \asymp_{d,p} (N \log N)^{-1/d},$$

and

$$\inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} (\widehat{W} - W_p(P, Q))^2 \asymp_{d,p} (N \log N)^{-2/d}.$$

Proof. The lower bounds are exactly Theorem 2.1(iv). The matching upper bounds follow from Theorem 12.8. \square

13 The boundary case $p = 1$: cone reduction and exact dyadic prefactors

At $p = 1$, the nearest-neighbor residuals are no longer semiconcave at the origin. The decisive point is that this singularity is only critical on the *normalized* local cube. In the exact dyadic decomposition every boundary term carries an additional global prefactor of order $h_J \asymp (N \log N)^{-1/d}$, and this prefactor makes the entire cone layer globally subcritical.

For $p = 1$ and dyadic level $J \geq 1$, define the piecewise linearized cost

$$c_{J,1}^{\text{lin}}(x, y) := \begin{cases} 0, & R = S, \\ \|c_R - c_S\|_2 + \frac{c_R - c_S}{\|c_R - c_S\|_2} \cdot [(x - c_R) - (y - c_S)], & R \neq S, \end{cases}$$

for $x \in R$, $y \in S$, and set

$$T_{J,1}^{\text{lin}}(P, Q) := \inf_{\pi \in \Pi(P, Q)} \int c_{J,1}^{\text{lin}}(x, y) d\pi(x, y).$$

Proposition 13.1 (Critical linearization and far-offset elimination at $p = 1$). *Assume $d > 2$, $n = m = N$, and choose the critical dyadic level J so that $2^{Jd} \asymp N \log N$. Then*

$$\sup_{P, Q \in \mathcal{P}_d} |W_1(P, Q) - T_{J,1}^{\text{lin}}(P, Q)| \lesssim_d (N \log N)^{-1/d}.$$

Moreover, same-cell contributions are diagonal at scale 2^{-J} , while every far offset with $\|m\|_\infty \geq 2$ is semiconcave after normalization and is therefore subcritical by Theorem 11.1.

Proof. The linearization error is the first-order Taylor remainder of the Euclidean norm on each dyadic macro-edge and is therefore $O(2^{-J})$ uniformly over the two cells. If the two points lie in the same cell, then their Euclidean distance is at most $\sqrt{d}2^{-J}$, so the same-cell contribution is diagonal at that scale. For a far offset m with $\|m\|_\infty \geq 2$, the translated difference body $m + Q_\square$ stays a positive distance away from the origin. Since the Hessian of $\|z\|_2$ is $O(\|z\|_2^{-1})$, multiplying the remainder by $\|m\|_2$ yields a uniformly bounded, uniformly Lipschitz, and uniformly semiconcave normalized family on Q_\square . Because $2^{-J} \asymp (N \log N)^{-1/d}$, the displayed bound follows. \square

Proposition 13.2 (Cone normal form). *Up to signed permutations and affine perturbations, every nearest-neighbor residual at $p = 1$ reduces to one of the d canonical cone kernels*

$$w \mapsto \sqrt{k} \|w\|_2, \quad w \in [0, 2]^k \times [-1, 1]^{d-k}, \quad k = 1, \dots, d.$$

Proof. Every nearest-neighbor offset can be sent by a signed permutation to $m^{(k)} := e_1 + \dots + e_k$ for some k . Writing the normalized residual in the coordinates $z = -m^{(k)} + w$ gives the exact identity

$$\Psi_{m^{(k)}}^{(1)}(-m^{(k)} + w) = \sqrt{k} \|w\|_2 - \sum_{j=1}^k w_j.$$

Removing the affine term leaves the stated cone kernel. \square

Theorem 13.3 (Finite cone switch envelopes are normalized-subcritical). *Assume $d > 2$. Let $K_X, K_Y \subset \mathbb{R}^d$ be fixed compact convex sets, let f_1, \dots, f_L be any finite family of translated cone kernels arising from Theorem 13.2 on $K_X \times K_Y$, let $\Lambda = [-M, M]^L$, and define*

$$g_\lambda := \min_{1 \leq j \leq L} (f_j - \lambda_j), \quad T_\lambda := T_{g_\lambda}, \quad \widehat{T}_{\lambda, N} := T_{g_\lambda}(\widehat{\mu}_N, \widehat{\nu}_N).$$

Then

$$\sup_{\lambda \in \Lambda} \sup_{\mu \in \mathcal{P}(K_X), \nu \in \mathcal{P}(K_Y)} \mathbb{E} |\widehat{T}_{\lambda, N} - T_\lambda(\mu, \nu)| \lesssim_{d, K_X, K_Y, L, M} N^{-1/d},$$

and

$$\sup_{\lambda \in \Lambda} \sup_{\mu \in \mathcal{P}(K_X), \nu \in \mathcal{P}(K_Y)} \mathbb{E} (\widehat{T}_{\lambda, N} - T_\lambda(\mu, \nu))^2 \lesssim_{d, K_X, K_Y, L, M} N^{-2/d}.$$

Proof. Write $a := 1/d \in (0, 1/2)$. Each translated cone kernel obeys the normalized empirical law $O(m^{-a})$ in absolute loss and $O(m^{-2a})$ in squared loss on the fixed compact convex class $K_X \times K_Y$.

Fix $\lambda \in \Lambda$, $\mu \in \mathcal{P}(K_X)$, and $\nu \in \mathcal{P}(K_Y)$. Exactly as in the proof of Theorem 12.7, decompose an optimal coupling for the envelope cost g_λ into active branches, label the empirical observations by the induced branch weights, thin to matched counts m_j , and use the resulting branchwise empirical plans as a feasible plan for $\widehat{T}_{\lambda, N}$. Because $(m_j/N)m_j^{-a} \leq N^{-a}$, $(m_j/N)^2 m_j^{-2a} \leq N^{-2a}$, and the leftover mass is again $O(N^{-1/2})$ in L^1 and $O(N^{-1})$ in L^2 , this gives the fixed- λ upper-bias bounds

$$\sup_{\lambda, \mu, \nu} \mathbb{E} (\widehat{T}_{\lambda, N} - T_\lambda(\mu, \nu))_+ \lesssim_{d, K_X, K_Y, L, M} N^{-a},$$

and

$$\sup_{\lambda, \mu, \nu} \mathbb{E} (\widehat{T}_{\lambda, N} - T_\lambda(\mu, \nu))_+^2 \lesssim_{d, K_X, K_Y, L, M} N^{-2a}.$$

For each fixed λ , reverse bias is again automatic by Theorem 12.5. To pass from fixed λ to the whole parameter cube, apply Theorem 12.6 with $\varepsilon := N^{-a}$. Because $a < 1/2$, the resulting centered fluctuation terms $N^{-1/2} \sqrt{\log N}$ and $N^{-1} \log N$ are both $o(N^{-a})$ and $o(N^{-2a})$, respectively. Combining the one-sided upper-bias estimate, the reverse-bias estimate, and the centered fluctuation bound yields the stated absolute and squared rates. \square

Lemma 13.4 (Dyadic prefactor restoration). *At the critical level $2^{Jd} \asymp N \log N$, every boundary cone block enters the exact dyadic decomposition with deterministic prefactor $h_J := 2^{-J} \asymp \eta_N$. Consequently,*

$$h_J N^{-1/d} = o(\eta_N), \quad h_J^2 N^{-2/d} = o(\eta_N^2).$$

Proof. Since $h_J \asymp \eta_N = (N \log N)^{-1/d}$, the two ratios are exactly $N^{-1/d} \rightarrow 0$ and $N^{-2/d} \rightarrow 0$. \square

Proposition 13.5 (Lifted closure of the boundary cone layer). *Assume $d > 2$ and $n = m = N$. Then there exists an estimator \widehat{W}_N such that*

$$\sup_{P, Q \in \mathcal{P}_d} \mathbb{E} |\widehat{W}_N - W_1(P, Q)| \lesssim_d \eta_N, \quad \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} (\widehat{W}_N - W_1(P, Q))^2 \lesssim_d \eta_N^2.$$

Proof. Let J be the critical level from Theorem 13.1, and write $h_J := 2^{-J}$. By Theorem 13.1, it is enough to estimate the linearized functional $T_{J,1}^{\text{lin}}(P, Q)$ at absolute scale η_N and squared scale η_N^2 .

The exact offset decomposition underlying Theorem 13.1 writes each nonzero macro-edge as a deterministic baseline $h_J \|m\|_2$, an affine term with global prefactor h_J , and a normalized residual term with global prefactor $h_J \|m\|_2^{-1}$.

Same-cell transport. If $m = 0$, then $R = S$, and therefore

$$W_1(P, Q) \leq \sqrt{d} h_J = O(\eta_N).$$

The same-cell part contributes $O(\eta_N)$ in absolute loss and $O(\eta_N^2)$ in squared loss.

Affine first-order part. After dividing by the global prefactor h_J , the local cost is affine on $Q_{\square} \times Q_{\square}$. Therefore Theorem 11.2 gives normalized fluctuation $O(N^{-1/2})$ in absolute loss and $O(N^{-1})$ in squared loss. After restoring the prefactor h_J , this contributes

$$O(h_J N^{-1/2}) = o(\eta_N)$$

in absolute loss and

$$O(h_J^2 N^{-1}) = o(\eta_N^2)$$

in squared loss.

Far offsets. If $\|m\|_{\infty} \geq 2$, Theorem 13.1 places the normalized residuals in one fixed compact semiconcave class. Hence Theorem 11.1 gives normalized empirical deviation $O(\rho_d(N))$ in absolute loss and $O(\rho_d(N)^2 + N^{-1})$ in squared loss. Since $h_J \|m\|_2^{-1} \leq h_J \asymp \eta_N$, the lifted far-offset contribution is again $o(\eta_N)$ and $o(\eta_N^2)$.

Terminal nearest-neighbor cone layer. By Theorems 13.1 and 13.2, the only unresolved terminal family is the finite cone-switch class on the nearest-neighbor macrograph. Its normalized empirical fluctuation is controlled by Theorem 13.3 at rate $N^{-1/d}$ in absolute loss and $N^{-2/d}$ in squared loss. By Theorem 13.4, restoring the exact dyadic prefactor yields the global bounds

$$O(h_J N^{-1/d}) = o(\eta_N), \quad O(h_J^2 N^{-2/d}) = o(\eta_N^2).$$

Combining the linearization error, the same-cell term, the lifted affine term, the lifted far-offset term, and the lifted terminal cone term yields the claimed estimator \widehat{W}_N . \square

Theorem 13.6 (Exact unrestricted balanced supercritical minimax law at $p = 1$). *Assume $d > 2$ and $n = m = N$. Then*

$$\inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} |\widehat{W} - W_1(P, Q)| \asymp_d (N \log N)^{-1/d},$$

and

$$\inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} (\widehat{W} - W_1(P, Q))^2 \asymp_d (N \log N)^{-2/d}.$$

Proof. The lower bounds are exactly Theorem 2.1(iv) at $p = 1$. The matching upper bounds are supplied by Theorem 13.5. \square

Theorem 13.7 (Exact unrestricted balanced supercritical minimax law for every $p \geq 1$). *Assume $p \geq 1$, $d > 2p$, and $n = m = N$. Then*

$$\inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}_d} \mathbb{E} |\widehat{W} - W_p(P, Q)| \asymp_{d,p} (N \log N)^{-1/d},$$

and

$$\inf_{\widehat{W}} \sup_{P, Q \in \mathcal{P}_d} \mathbb{E}(\widehat{W} - W_p(P, Q))^2 \asymp_{d,p} (N \log N)^{-2/d}.$$

Proof. Combine Theorems 12.3, 12.9 and 13.6. \square

14 Arbitrary sample sizes and the complete fixed-dimensional picture

The known-reference lower bound from Theorem 10.2 shows that the supercritical hard family already survives when one marginal is known exactly. Once the balanced closure Theorem 13.7 is available, sample-size imbalance therefore plays no further role.

Theorem 14.1 (Exact unrestricted supercritical minimax law for arbitrary sample sizes). *Assume $p \geq 1$, $d > 2p$, and $N := n \wedge m \geq 2$. Then*

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

Proof. For the upper bounds, discard all but the first N observations from the larger sample and apply Theorem 13.7 to the resulting balanced experiment. For the lower bounds, first assume $n = N \leq m$ and fix the reference measure Q_N from Theorem 10.2. Given any two-sample estimator \widehat{W} , average it over an independent m -sample from Q_N to obtain a known-reference estimator. Jensen's inequality shows that this averaging cannot decrease the restricted risk on the submodel $Q = Q_N$. The one-sided lower bound of Theorem 10.2 therefore yields the same $(N \log N)^{-1/d}$ and $(N \log N)^{-2/d}$ lower bounds. The case $m = N \leq n$ is symmetric. \square

Theorem 14.2 (Full fixed-dimensional minimax picture). *Fix $d \geq 1$, $p \geq 1$, and $N := n \wedge m \geq 2$.*

(i) *If $d < 2p$, then*

$$\mathfrak{M}_{n,m,d,p}^{\text{abs}} \asymp N^{-1/2}, \quad \mathfrak{M}_{n,m,d,p}^{\text{sq}} \asymp N^{-1}.$$

(ii) *If $d = 2p$, then*

$$N^{-1/2} \lesssim \mathfrak{M}_{n,m,d,p}^{\text{abs}} \lesssim N^{-1/2} (\log N)^{1/p},$$

$$N^{-1} \lesssim \mathfrak{M}_{n,m,d,p}^{\text{sq}} \lesssim N^{-1} (\log N)^{2/p}.$$

(iii) *If $d > 2p$, then*

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

Proof. Parts (i) and (ii) are exactly Theorem 2.1(i,ii). Part (iii) is Theorem 14.1. \square

Remark 14.3. In the supercritical regime, extra observations from one side are globally useless at minimax level. Once the smaller sample size $N = n \wedge m$ is fixed, enlarging the other sample cannot change the order of the unrestricted risk. Thus the two-sample minimax law depends on (n, m) only through N .

Conclusion

The unrestricted fixed-dimensional supercritical minimax problem for two-sample Wasserstein distance estimation is closed here through a fully Euclidean chain of arguments. The sharp lower-envelope scale first appears on a full diagonal neighborhood, the balanced problem is identified with its critical dyadic-grid model, the generic off-diagonal geometry is resolved by positive-baseline and mixed-scale theorems, and the only terminal obstruction that survives the reduction is a finite nearest-neighbor layer. That terminal layer is then closed by semiconcavity for $p \geq 2$, by corner kernels and finite switch envelopes for $1 < p < 2$, and by exact prefactor restoration for $p = 1$.

Methodologically, the paper isolates three ingredients that appear to be decisive for the scalar Wasserstein functional: a genuinely Euclidean diagonal theorem at the lower-envelope scale, an exact critical-grid model with $\asymp N \log N$ atoms, and a final singular analysis in which the correct dyadic prefactors are restored before passing from the transport cost W_p^p back to the distance W_p . Taken together, these ingredients show that the logarithmic enlargement from N to $N \log N$ is neither a tree artefact nor a purely diagonal curiosity. It is an intrinsic Euclidean phenomenon of the scalar functional $W_p(P, Q)$, visible only after the multiscale and singular bookkeeping are carried through to completion.

References

- [1] 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.
- [2] 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.
- [3] J. Jiao, Y. Han, and T. Weissman. Minimax estimation of the L_1 distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018.
- [4] E. M. Bronshtein. ε -entropy of convex sets and functions. *Siberian Mathematical Journal*, 17:393–398, 1976.
- [5] A. Guntuboyina and B. Sen. L_1 covering numbers for uniformly bounded convex functions. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *JMLR Workshop and Conference Proceedings*, pages 12.1–12.13, 2012.
- [6] 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.
- [7] J. Niles-Weed and P. Rigollet. Estimation of Wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022.
- [8] S. Wang, T. T. Cai, and H. Li. Optimal estimation of Wasserstein distance on a tree with an application to microbiome studies. *Journal of the American Statistical Association*, 116(535):1237–1253, 2021.
- [9] 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.