TUNED CONVEX AUGMENTATION BEATS THE SAMPLE IT IS BUILT FROM

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-augmentation methods such as SMOTE and mixup create synthetic points as random convex combinations of K real examples. We ask when the resulting synthetic distribution is closer to the data distribution, in Wasserstein distance, than the n real examples it is built from, whose empirical distribution converges at rate $n^{-1/d}$. Mixing nearby points smooths the sample. This reduces fluctuation at the cost of a bias of order h^2, where h is the neighbourhood size. For a smooth density bounded away from zero, we prove that k-NN SMOTE with tuned k beats $n^{-1/d}$ for every $K \geq 2$ in every dimension $d \geq 3$. Its rate reaches $n^{-2/(d+2)}$ once $K \geq d$, and for a non-flat density on the torus no local mixing scheme with a bias of order h^2 can do better. The gain carries over to a finite training set only if synthetic points make up most of it. In experiments, tuned mixing of pairs beats the real sample for n as small as 90. In full dimension a tuned kernel does better, while on the two manifolds we test mixing beats an isotropic kernel at every sample size and ambient dimension.

1 INTRODUCTION

Many data-augmentation methods create new training points as random convex combinations of existing ones. mixup (Zhang et al., 2018b) interpolates between two arbitrary training points and their labels. SMOTE (Chawla et al., 2002) interpolates between a minority-class sample and one of its k nearest neighbours, and simplicial SMOTE (Kachan et al., 2025) does the same across a simplex of mutually neighbouring points. Manifold mixup (Verma et al., 2019), token- and patch-level variants (Choi et al., 2022; Bae et al., 2023), latent mixup (Kang et al., 2022) and clustering-aware mixup (Xie et al., 2023) apply the same operation in other representations, and Local Mixup (Baena et al., 2022) and k-Mixup (Greenewald et al., 2023) favour nearby points, which is the locality that our bounds require. Empirically, these methods improve accuracy, calibration and robustness (Thulasidasan et al., 2019; Zhang et al., 2022; Pinto et al., 2022). Theoretical accounts usually treat them as vicinal risk minimisation (Zhang et al., 2018a; Mangla et al., 2021), as implicit regularisation (Carratino et al., 2022; Guo et al., 2019; Zhang et al., 2021) or as interpolation (Shimada et al., 2019; Kimura, 2021).

Much less is known about the synthetic sample itself as a statistical object. For SMOTE, existing theory describes the distribution of a single synthetic point (Elreedy & Atiya, 2019; Elreedy et al., 2024), its behaviour in high dimension (Blagus & Lusa, 2013), its tendency to copy the minority sample (Sakho et al., 2024), its consistency (Kamalov et al., 2026) and when oversampling helps (Dablain & Chawla, 2023; Stocksieker et al., 2023). None of these results gives a rate for the synthetic distribution, and Kamalov et al. (2026) recommend $k = 1$, whereas in our analysis k grows with n. Concurrently, Xia & Klusowski (2026) bound a SMOTE cost of order $(k/N_1)^{1/d}$, where N_1 is the minority count, inside a total-variation bound on the risk. The corresponding quantity in our analysis is the mixing bias in W_1, which is of order $(k/N_1)^{2/d}$ for smooth μ_0. A generator that produces many synthetic points from n real ones reuses each real point in many combinations, so the synthetic sample carries no more information than the n real points, and this reuse enters our W_1 bounds only through Kish's effective sample size (Kish, 1965).

The natural benchmark is the empirical measure, whose W_1 rate $n^{-1/d}$ is sharp for $d \geq 3$ (Dudley, 1969; Fournier & Guillin, 2015; Boissard & Le Gouic, 2014; Weed & Bach, 2019). Smoothing can do better. Kernel estimators attain the minimax rate $n^{-(1+s)/(d+2s)}$ for s-smooth densities (Weed & Berthet, 2019; Niles-Weed & Berthet, 2022), and they do so also on manifolds (Divol, 2022).

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

Gaussian smoothing speeds up the convergence of empirical measures (Goldfeld et al., 2020; Chen & Niles-Weed, 2022). Tuned mixing can be read as a smoothed bootstrap (Silverman & Young, 1987) whose kernel is built from the data, and these fixed-kernel methods are its counterparts. Rates are also known for generative models such as GANs (Liang, 2021) and diffusion models (Oko et al., 2023), but they concern learned samplers rather than a fixed convex operation. Convex mixing is not a kernel estimator in the usual sense. Its kernel is built from the data, the same points are reused in many combinations, and each combination lives on a $(K-1)$-dimensional simplex. It is therefore not clear in advance whether it can beat the sample it is built from.

We show that it can. When k is chosen by a simple rule, k-NN SMOTE attains the rate $n^{-2/(2d-\bar{\kappa})}$, where $\bar{\kappa} = \min(K-1, d-2)$, and at $K = d-1$ this rate carries an extra logarithmic factor. This beats $n^{-1/d}$ for every arity when $d \geq 3$. For a non-flat density on the torus, no convex mixing of locality h with a bias of order h^2 beats $n^{-2/(d+2)}$. The gain belongs to the synthetic distribution, and a training set of n synthetic points does not keep it. On the torus a tuned kernel smoother attains the same rate, and which of the two wins is a matter of constants and geometry. We study local mixing within one class, as SMOTE does, and our results also apply to Local Mixup and k-Mixup within each class. Our aim is to quantify what this mixing delivers, and we do not propose it as a density estimator. Global mixup, which mixes arbitrary pairs and their labels, estimates a different distribution. In more detail, we make four contributions.

Upper and lower bounds. We prove the rate and a lower bound that matches it from $K = d-1$ on (Theorems A and B). A combination of K points smooths in at most $K-1$ directions, and the rate improves with this number until it reaches $d-2$, after which the fluctuation is already that of a full-dimensional kernel (Section 3). For $K \geq d$ the same rate holds in W_2 on the torus (Theorem F.1).

The price of smoothing. We compute the second-order bias for Dirichlet(α) weights, for a fixed radius and for the k-NN rule, and bound a boundary term of the same order (Proposition 4.1). Most of this bias comes from choosing a neighbour in a density gradient rather than from interpolating between points. The computed interior constant agrees with simulations without any fitted parameter.

What the rate means in practice. We give the number and weight of synthetic points that a training set needs to keep the gain (Corollary 3.1). Our bound gives no gain unless $k > C_Q^d$, and this threshold does not depend on n. If C_Q is about 1.3, which our measurements allow, SMOTE's default $k = 5$ falls below it from about $d = 6$ (Section 5.3).

Mixing versus kernel smoothing. At a boundary, the bias of a kernel has the same order as that of mixing (Lemma D.5). On a manifold, mixing leaves the manifold only at second order in the locality, whereas a Euclidean kernel leaves it at first order in the bandwidth (Proposition 5.1). In experiments (Section 6), tuned mixing is 21–35% below the sample in every population from $n = 90$ on, whereas SMOTE's default $k = 5$ gains 6–15%. On two manifolds it beats an isotropic kernel by 13–34% at every n and codimension. In full dimension, a tuned kernel has the better constant. When k is tuned for each arity, no larger arity beats pairs by more than 2% in the bounded domain and on the two manifolds, as the bias constants of Section 4 suggest.

2 Setup

Let $X_1, \ldots, X_n$ be i.i.d. from μ_0 on $\mathbb{R}^d$. Our main example is k-NN SMOTE. To keep the bases independent of the neighbourhoods they are completed with, we analyse it with sample splitting.

Definition 2.1 (k-NN SMOTE with splitting)**.** Split the sample into a pool $\mathcal{D}_A$ of $n_A = \lfloor n/2 \rfloor$ points and $n_B = n - n_A$ bases $\mathcal{D}_B$. A synthetic point is $Y = U_1 X_b + \sum_{j=2}^{K} U_j Z_j$, where X_b is uniform on $\mathcal{D}_B$, $\{Z_2, \ldots, Z_K\}$ is a uniform set of $K-1$ of the k nearest neighbours of X_b in $\mathcal{D}_A$ with $k \geq K-1$, and $U \sim \mathrm{Dir}(\alpha, \ldots, \alpha)$ with $\alpha > 0$ is drawn independently. The distribution of Y given the sample is $\widehat{\nu}_X = \frac{1}{n_B} \sum_{b \in \mathcal{D}_B} \Phi_{X_b}$, where Φ_x is the distribution of Y given $X_b = x$, and we call it the *synthetic distribution*. A *synthetic sample* consists of N independent draws from it.

Each base is used equally often, and $\alpha = 1$ gives SMOTE's flat weights. The *locality* $h_\star$ is the largest distance between two points of a combined tuple. For k-NN SMOTE it is of order $D(k/n)^{1/d}$. A tuple of K points spans a simplex of dimension at most $K-1$, and mixing smooths within it. Our proofs control the smoothing in $\kappa := \min(K-1, d-1)$ directions, and the rates depend on the arity

Table 1: Rate exponents of the empirical measure, Theorem A(i) and Proposition E.10. Bold: $2/(d+2)$.

rate exponent	$d=3$	4	6	8	10	rule
empirical measure, $K=1$	.333	.250	.167	.125	.100	$1/d$
$K=2$ (pairs)	**.400**	.286	.182	.133	.105	$2/(2d-1)$
$K \geq d-1$	**.400**	**.333**	**.250**	**.200**	**.167**	$2/(d+2)$
$K=2$ under Assumption 3	**.400**	**.333**	.200	.143	.111	$1/(d-1)$, $d \geq 4$

through $\bar{\kappa} := \min(K-1, d-2)$. The rate theorem is stated for $\alpha = 1$, while the bias expansion and the lower bound hold for every $\alpha > 0$. We assume the following.

Assumption 1 (Population)**.** μ_0 is carried by the closure of a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ of diameter D and has a density $f \in C^2(\bar{\Omega})$ with $f \geq f_{\min} > 0$ on Ω. Then $\mu_0(B(x,r)) \geq c_0 r^d$ for $x \in \Omega$, $r \leq D$, and we put $c_f := \omega_d \|f\|_\infty / c_0 \geq 1$.

The dependence between synthetic points enters our bounds through a single number. Let w_i be the share of mixing slots filled by X_i. Then $S = \sum_i w_i^2 \geq 1/n$ is the inverse of Kish's effective sample size (Kish, 1965), and all fluctuation bounds hold with $1/n$ replaced by S (Section A). For SMOTE with splitting, $S = 1/n_A$.

3 Main results

Our upper bounds compare the synthetic distribution with μ_0 through the triangle inequality

$$W_p(\widehat{\nu}_X, \mu_0) \;\leq\; \underbrace{W_p(\widehat{\nu}_X, \bar{\nu})}_{\text{fluctuation}} \;+\; \underbrace{W_p(\bar{\nu}, \mu_0)}_{\text{mixing bias}}, \qquad \bar{\nu} := \mathbb{E}\widehat{\nu}_X, \tag{1}$$

which we use mostly with $p = 1$, and we write $\beta_1 := W_1(\bar{\nu}, \mu_0)$. The mixing bias is deterministic, and all the randomness is in the fluctuation. A smaller locality reduces the bias, but the synthetic distribution is then smoothed less and its fluctuation is larger. Choosing the locality balances the two terms, and our first result gives the rate of k-NN SMOTE when k is chosen in this way.

Theorem A (Tuned convex augmentation beats its sample)**.** *Let $\widehat{\nu}_X$ be the synthetic distribution of Definition 2.1 with $\alpha = 1$, built from i.i.d. data from μ_0 satisfying Assumption 1 with $d \geq 3$. Let $K \geq 2$, $\bar{\kappa} := \min(K-1, d-2)$ and*

$$k = \left\lceil n^{(d-\bar{\kappa})/(2d-\bar{\kappa})} \right\rceil. \tag{2}$$

There are constants C and n_0, which depend only on d, K, Ω and f, such that for $n \geq n_0$

(i) $\mathbb{E}\, W_1(\widehat{\nu}_X, \mu_0) \leq C\, n^{-2/(2d-\bar{\kappa})}\, \ell_n$, *where $\ell_n = \log n$ if $K = d-1$ and $\ell_n = 1$ otherwise;*

(ii) *for every $\delta \in (0,1)$, with probability at least $1 - \delta - ne^{-k/2}$,*

$$\left| W_1(\widehat{\nu}_X, \mu_0) - \mathbb{E} W_1(\widehat{\nu}_X, \mu_0) \right| \;\leq\; 21(K-1)\, c_f\, D \sqrt{\log(2/\delta)/n}. \tag{3}$$

The rule (2) equates the two terms of (1). The mixing bias is of order $h_\star^2 \asymp D^2 (k/n)^{2/d}$, so it grows with k. The fluctuation decreases as k grows and the synthetic distribution becomes smoother, and it is of order $D\, n^{-1/(d-\bar{\kappa})} (h_\star / D)^{-\bar{\kappa}/(d-\bar{\kappa})}$, apart from a logarithmic factor at $K = d-1$. The two terms are equal when $h_\star \asymp D n^{-1/(2d-\bar{\kappa})}$, which is the rule (2), and both are then of order $n^{-2/(2d-\bar{\kappa})}$. The arity enters only through the exponent $\bar{\kappa}$ of the fluctuation, which the chaining argument below explains. The deviation in (ii) is of order $n^{-1/2}$, which is smaller than the rate.

Theorem A′ in Section E extends Theorem A to the general reuse model, to every k and to the ε-ball rule, which takes as partners the pool points within distance $\varepsilon \propto n^{-1/(2d-\bar{\kappa})}$ of the base. In W_2 the same generator attains $n^{-2/(d+2)}$ on the torus for $K \geq d$, while with a boundary the tuned W_2 rate is open (Section F).

Since the empirical measure $\hat{\mu}_n$ of the sample has $\mathbb{E} W_1(\hat{\mu}_n, \mu_0) \geq c n^{-1/d}$, Theorem A shows that tuned SMOTE beats the sample it is built from. In the setting of Theorem A,

$$\frac{\mathbb{E} W_1(\widehat{\nu}_X, \mu_0)}{\mathbb{E} W_1(\hat{\mu}_n, \mu_0)} \;\leq\; C' n^{-\bar{\kappa}/(d(2d-\bar{\kappa}))}\, \ell_n \;\longrightarrow\; 0, \tag{4}$$

where C' depends only on d, K, Ω and f. The exponent of Theorem A increases with K and saturates at $2/(d+2)$ from $K = d-1$ on (Table 1), while the same tuning at $K = 1$ returns the sample's exponent $1/d$. For $d \geq 4$ the proved exponents therefore favour arities above two, which is what simplicial SMOTE (Kachan et al., 2025) uses when it mixes $K > 2$ mutually neighbouring points.

The exponent in Theorem A comes from a chaining argument. We sketch this argument because it shows where the arity enters and why the threshold $\kappa = d - 2$ appears.

Why the empirical measure stops at $n^{-1/d}$. Partition the support into dyadic cubes of relative side $t_\ell = 2^{-\ell}$ for $\ell \leq L$. Moving mass first between the children of each cube and then within the cubes of the finest level gives, for any two probability measures (Section A),

$$W_1(\nu, \nu') \;\leq\; \sqrt{d}\, D\Big(t_L + 2\sum_{\ell=1}^{L} t_\ell A_\ell\Big), \qquad A_\ell = \sum_{Q \in \mathcal{Q}_\ell} |\nu(Q) - \nu'(Q)|. \tag{5}$$

For $\nu = \widehat{\nu}_X$ and $\nu' = \bar{\nu}$, the Cauchy–Schwarz inequality over the t_ℓ^{-d} cubes gives $\mathbb{E}A_\ell \leq t_\ell^{-d/2}(\sum_Q \operatorname{Var}\widehat{\nu}_X(Q))^{1/2}$. For the empirical measure $\sum_Q \operatorname{Var}\hat{\mu}_n(Q) \leq 1/n$, so level ℓ costs $n^{-1/2}t_\ell^{1-d/2}$. When $d \geq 3$ this cost grows as the cubes shrink, while the finest-level term t_L decreases. The two balance at the *sampling scale* $t_L \asymp n^{-1/d}$, where a cube contains about one sample point, and this gives the empirical rate.

What mixing changes. A synthetic point is a random point of a simplex of dimension κ and diameter $h_\star$. Conditionally on the pool, the mass of a cube of side s fluctuates only to the extent that two points of the same simplex tend to fall in the same cube, and the randomness of the pool adds the probability that a single point falls in a given cube. Both are small-ball probabilities. Two independent points of a κ-dimensional set of diameter $h_\star$ are within distance s of each other with probability of order $(s/h_\star)^\kappa$, and accordingly, for the k-NN generator truncated at a radius that holds its neighbours with high probability,

$$\sum_Q \operatorname{Var}\widehat{\nu}_X(Q) \;\leq\; \frac{C}{n}\Big(\frac{s}{h_\star} \wedge 1\Big)^\kappa, \tag{6}$$

where C depends on d, K and f (Lemmas E.2 to E.4). The bound holds only on average over the random simplex, because nearly flat simplices occur with positive probability. It also sees one simplex at a time. Two points from one base share a partner only with probability of order $1/k$, and Lemma E.9 raises the leading exponent for the bases to $\min\{2(K-1), d-1\}$. If the pool obeys the same bound (Assumption 3), as Monte Carlo indicates, every $K > d/2$ attains $n^{-2/(d+2)}$ and pairs gain the last row of Table 1 (Proposition E.10).

Running the chain below the sampling scale. With (6), a level below the smoothing scale $t_{\mathrm{sm}} := h_\star/(C_Q D)$, where $C_Q := \sqrt{d}$ is the diameter of the unit cube, costs

$$n^{-1/2}\, t_{\mathrm{sm}}^{-\kappa/2}\; t_\ell^{(\kappa-(d-2))/2}, \tag{7}$$

instead of $n^{-1/2}t_\ell^{1-d/2}$, so the chain can be continued past the sampling scale.

First, smoothing helps only if $t_{\mathrm{sm}} > n^{-1/d}$, because blurring below the statistical resolution has no effect at any arity (Proposition B.5). For k-NN locality $h_\star \asymp D(k/n)^{1/d}$, so the condition reads $k > C_Q^d$ up to the constant of the k-NN radius, and it does not involve n (Section 5.3).

Second, the sign of $\kappa - (d-2)$ decides how much of the gain survives (Proposition B.4). If $\kappa > d-2$, the costs (7) decrease towards the fine levels, and their sum is of the order of the coarsest one. The fluctuation is then $D\, n^{-1/2}(h_\star/D)^{1-d/2}$, which is the fluctuation of a d-dimensional kernel of width $h_\star$, and balancing it against a bias of order $h_\star^2$ gives $n^{-2/(d+2)}$. If $\kappa < d-2$, the costs grow towards the finest level. Balancing that level against t_L leaves $D\, n^{-1/(d-\kappa)}(h_\star/D)^{-\kappa/(d-\kappa)}$, which gives the rows of Table 1 below $K = d-1$. At $\kappa = d-2$ a logarithm appears.

The threshold $\kappa = d-2$ is not an artefact of chaining. A κ-dimensional measure has finite Newtonian ($\dot{H}^{-1}$) energy if and only if $\kappa > d-2$, and in that range Peyre's inequality gives the same fluctuation

on the torus without chaining. Below the threshold, a covering argument shows that the loss is real when about n simplices are drawn. The general statement is Theorem B.3 (Sections B and C).

The chaining argument thus explains why the rate of Theorem A stops improving at $n^{-2/(d+2)}$ once $K \geq d-1$. The next theorem shows that no local convex mixing scheme does better on the torus when the density is not flat, whatever arity and weights it uses, as long as its bias is at least of order h^2.

Theorem B (Lower bound for local convex mixing)**.** *Let $d \geq 3$ and let μ_0 on $\mathbb{T}^d$ have a density $f \in C^2$ with $f \geq f_{\min} > 0$ and $\nabla f \not\equiv 0$. Let $\widehat{\nu}_X = n_B^{-1} \sum_{b \in \mathcal{D}_B} \Phi_{X_b}$, where given $\mathcal{D}_A$ the bases X_b are i.i.d. μ_0 and each distribution Φ_x is carried by $B(x,h)$, and suppose that its mixing bias $\beta_1 := W_1(\mathbb{E}\widehat{\nu}_X, \mu_0)$ obeys $\beta_1 \geq c_\beta h^2$ for $h \leq h_0$. Then there are $c > 0$ and n_0 such that for every $h \in (0, h_0]$ and $n \geq n_0$,*

$$\mathbb{E}W_1(\widehat{\nu}_X, \mu_0) \;\geq\; c \max\left\{h^2,\, \min\{n^{-1/2}h^{1-d/2},\, n^{-1/d}\}\right\} \;\geq\; c\, n^{-2/(d+2)}.$$

The bias hypothesis holds for the fixed-radius rule of Proposition 4.1 and for the ε-ball rule, at every $K \geq 2$ and for every $\mathrm{Dir}(\alpha, \ldots, \alpha)$ weights, and for k-NN SMOTE truncated at the radius $h = (4k/(c_0 n_A))^{1/d}$, which holds the k nearest neighbours with high probability, by Lemma E.7.

The two terms of the lower bound have different origins. The bias term is the hypothesis, which Jensen's inequality carries over to the synthetic distribution. The rules above satisfy it because non-negative weights with a partner radius that does not grow with the density act as a second-order kernel (Section 4). The fluctuation term uses only that each synthetic point stays within distance h of its base, so it holds for every weight distribution and does not depend on κ. From $K = d-1$ on, Theorem A, which also holds on the torus, attains the bound, apart from a logarithmic factor at $K = d-1$. Tuned k-NN SMOTE is therefore rate-optimal among local convex-mixing generators with a bias of order h^2, and this class contains the fixed-radius, ε-ball and k-NN rules at every arity and weight distribution. Below $K = d-1$ the true exponent lies between the one we prove, which is sharp for a generator that draws about n distinct tuples, and $2/(d+2)$ (Remark C.11).

In practice one trains on finitely many points drawn from the synthetic distribution $\widehat{\nu}_X$, usually together with the real ones, and the rate of Theorem A survives only if the synthetic points are numerous and carry most of the weight.

Corollary 3.1 (A synthetic training set)**.** *Let $\widehat{\nu}_{X,N}$ be the empirical measure of N points drawn from $\widehat{\nu}_X$ given the data, and let $M_\lambda := (1-\lambda)\hat{\mu}_n + \lambda\widehat{\nu}_{X,N}$ be a training measure that keeps the real points with total weight $1-\lambda$. For every realisation,*

$$W_1(M_\lambda, \mu_0) \;\geq\; c_{\mathrm{q}} \max\left\{(1-\lambda)^{1+1/d} n^{-1/d},\, (n+N)^{-1/d}\right\}, \qquad c_{\mathrm{q}} := \tfrac{1}{2}(2\|f\|_\infty \omega_d)^{-1/d}. \tag{8}$$

For $d \geq 3$ there is a constant C_d, which depends on d only, such that

$$\mathbb{E}W_1(M_\lambda, \mu_0) \;\leq\; (1-\lambda)\, C_d D n^{-1/d} + \mathbb{E}W_1(\widehat{\nu}_X, \mu_0) + C_d D N^{-1/d}. \tag{9}$$

At the parameters of Theorem A, M_λ therefore keeps the rate if $N \gtrsim n^{2d/(2d-\bar{\kappa})}$ and $1-\lambda \lesssim n^{-\bar{\kappa}/(d(2d-\bar{\kappa}))}$. It cannot beat $n^{-1/d}$ if $N \asymp n$ or if $1-\lambda$ stays bounded below, which is the case when SMOTE is run to class parity at a bounded imbalance ratio.

In particular, SMOTE run to class parity at imbalance ratio ρ keeps the rate if and only if $\rho \gtrsim n^{\bar{\kappa}/(2d-\bar{\kappa})}$, where n is the minority count.

Since W_1 controls expectations of Lipschitz functions, the rate also bounds the excess risk.

Corollary 3.2 (Excess risk of training on synthetic data)**.** *Let $\{\ell(\cdot,\theta)\}_{\theta\in\Theta}$ be L-Lipschitz in x uniformly in θ, and let $\hat{\theta}$ minimise the risk under $\widehat{\nu}_X$. At the parameters of Theorem A, with probability at least $1 - \delta - ne^{-k/2}$,*

$$R_{\mu_0}(\hat{\theta}) - \inf_{\theta\in\Theta} R_{\mu_0}(\theta) \;\leq\; 2L\Big(C\, n^{-2/(2d-\bar{\kappa})}\ell_n + 21(K-1)\, c_f D\sqrt{\log(2/\delta)/n}\Big). \tag{10}$$

Total variation cannot give this, since $d_{TV}(\widehat{\nu}_X, \mu_0) = 1$ for $K \leq d$ (Xia & Klusowski, 2026).

Proof ideas. The proofs combine a few ingredients. Because $\bar{\nu} = \mathbb{E}\widehat{\nu}_X$ is deterministic, Kantorovich–Rubinstein duality writes $W_1(\widehat{\nu}_X, \bar{\nu})$ as the supremum of a centred process, and there is no interpolation term. The small-ball estimate (6) is proved for k-NN tuples by peeling the simplex volume with Gram–Schmidt, including degenerate tuples, and for $K > d$ by aggregating the Dirichlet weights onto d of the points. Sample splitting makes the bases independent given the pool, and an Efron–Stein bound handles the randomness of the pool. The bias follows from a Kramers–Moyal expansion and the Beckmann form of W_1. For the lower bound, Jensen's inequality gives $\mathbb{E}W_1 \geq \beta_1$, and tent functions of width $C\max\{h, n^{-1/d}\}$, whose signs are chosen from the data, bound the fluctuation from below conditionally on the pool. The proofs are in Sections C and E.

4 The mixing bias

The smoothing that reduces the fluctuation also creates the bias term in (1). For a smooth density it is of second order in the locality, and in the interior of the domain its constant is explicit.

Proposition 4.1 (Second-order mixing bias)**.** *Let μ_0 satisfy Assumption 1, let $Z_1 \sim \mu_0$, let $Z_2, \dots, Z_K$ be i.i.d. from μ_0 restricted to $B(Z_1, h)$, and let $U \sim \mathrm{Dir}(\alpha, \dots, \alpha)$ be independent of them. Let $\bar{\nu}_h$ be the distribution of $Y = \sum_j U_j Z_j$ and let $\beta_1 := W_1(\bar{\nu}_h, \mu_0)$. Then*

$$\beta_1 \;\leq\; \big(c_K(\alpha)\,\|\nabla f\|_{L^1} + C_\partial\big)h^2 + o(h^2), \qquad c_K(\alpha) := \frac{(K-1)\big((2K-1)\alpha+1\big)}{2K(K\alpha+1)(d+2)}, \tag{11}$$

where $C_\partial \leq c_{d,K} f_{\max} \mathrm{Per}(\Omega)$ comes from the layer of width h along $\partial\Omega$ and vanishes on a torus. On the circle, if f' has median zero,

$$\beta_1 \;=\; c_K(\alpha)\, h^2\, \|f'\|_{L^1} + o(h^2). \tag{12}$$

For the flat weights of SMOTE, $c_K(1) = \frac{K-1}{(K+1)(d+2)}$.

Here h is the radius of the partner ball, so a tuple has spread at most $2h$. The first-order displacement cancels because the weights are exchangeable, which makes mixing a symmetric smoothing. A Kramers–Moyal expansion then gives $\bar{\nu}_h - \mu_0 = -\nabla \cdot V_0 + o(h^2)$ for the flux $V_0 = c_K(\alpha)h^2\nabla f$, and the Beckmann form of W_1 bounds β_1 by the L^1 norm of this flux. The proofs are in Sections D and E, and Section 6 compares the closed form with simulations.

Selection versus interpolation. As $\alpha \to 0$ the output becomes a uniformly chosen vertex of the tuple and there is no interpolation, so $c_K(0) = \frac{K-1}{2K(d+2)}$ is the bias caused by choosing a neighbour in a density gradient. For a fixed radius this is already a fraction $c_K(0)/c_K(1) = (K+1)/(2K)$ of the flat-weight bias. Since c_K increases with α, the small α that is common in mixup removes at most a fraction $(K-1)/(2K)$ of the bias. With the k-NN radius, which adapts to the density, the share of selection is larger still. It equals $(d+2)(K+1)/(2(dK+2))$, and small α also weakens the smoothing.

Why the bias limits the rate. Convex weights are non-negative, so mixing is a positive kernel. With a partner radius that does not grow with the density, its second moment does not vanish, and for a non-flat density $\beta_1 = \Theta(h^2)$. Theorem B turns this into the lower bound $n^{-2/(d+2)}$. This is the $s = 1$ case of the minimax rate $n^{-(1+s)/(d+2s)}$ for s-smooth densities (Weed & Berthet, 2019; Niles-Weed & Berthet, 2022). Over the C^2 class of Assumption 1 the minimax rate is $n^{-3/(d+4)}$, which is faster. A fixed-width kernel attains it only with negative weights, which convex mixing cannot produce. For k-NN SMOTE the radius is larger where f is small, and this multiplies the flat-weight constant by $1 + 2/(dK)$. A radius that grows with the density acts the other way. At a radius proportional to $f^{K/2}$ the interior h^2 term cancels, as it does for the kernel of Abramson's square-root law (Abramson, 1982), and such rules fall outside Theorem B.

The boundary. Near $\partial\Omega$ the partner ball is truncated, and the layer of width h along the boundary adds a term $C_\partial h^2$ of the same order as the interior one. We bound it but do not compute it, so the explicit constant is the full constant only on a torus. A Gaussian kernel also loses mass across $\partial\Omega$, which costs order σ^2 in W_1, so at the boundary mixing and kernel smoothing differ in their constants but not in the order of the bias.

5 MIXING VERSUS KERNEL SMOOTHING

Mixing is a form of smoothing, and Table 25 in Section J compares it with kernel estimators.

5.1 IN THE SETTING OF THEOREM A

The smoothed bootstrap $\hat{\mu}_n * \mathcal{N}_\sigma$, where $\mathcal{N}_\sigma$ is the centred Gaussian of covariance $\sigma^2 I_d$, attains the same rate $n^{-2/(d+2)}$ on the torus when $\sigma \asymp n^{-1/(d+2)}$ (Lemma D.5). Its bias is $O(\sigma^2)$, and its fluctuation behaves like (7) with $\kappa = d$. Neither the rate nor the limit $n^{-2/(d+2)}$ of an h^2 bias separates the two methods, and the differences lie in constants and in structure. A Gaussian kernel is non-degenerate, since it smooths in all d directions at every bandwidth, so no condition on κ is needed. Mixing, on the other hand, never leaves the convex hull of the sample. Convex constraints on the features, such as bounds, non-negativity or compositional constraints, are therefore preserved exactly, whereas a kernel moves mass across $\partial\Omega$ at a cost of order σ^2 in W_1. The convex combination commutes with affine maps of the features, and with k-NN locality the scale is set by k rather than by a bandwidth, so a common rescaling of the features changes nothing. Global mixing, whose locality is $h_\star = D$, does not smooth μ_0 at all. It estimates instead the distribution of a random convex combination of K independent draws from μ_0 (Corollary A.13).

5.2 ON A MANIFOLD

Proposition 5.1 (Mixing stays near a manifold, a Euclidean kernel does not)**.** *Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact C^2 submanifold of dimension $m < d$ and reach $\tau > 0$, and let μ_0 be carried by $\mathcal{M}$.*

(i) *If $Z_1, \dots, Z_K \in \mathcal{M}$ and $\max_j \|Z_j - Z_1\| \le h$, every convex combination Y of them has $\operatorname{dist}(Y, \mathcal{M}) \le h^2/\tau$ for any choice of the weights. Every h-local generator on data from μ_0 therefore has $\int \operatorname{dist}(\cdot, \mathcal{M})\, d\widehat{\nu}_X \le h^2/\tau$.*

(ii) *Let $\mathcal{N}_\sigma := \mathcal{N}(0, \sigma^2 I_d)$. For every probability ρ on $\mathcal{M}$ and $0 < \sigma \le \tau\sqrt{d-m}/(6d)$, $\int \operatorname{dist}(\cdot, \mathcal{M})\, d(\rho * \mathcal{N}_\sigma) \ge \sigma\sqrt{d-m}/5$, and hence $W_1(\rho * \mathcal{N}_\sigma, \mu_0) \ge \sigma\sqrt{d-m}/5$.*

(iii) *If moreover $\mu_0(B(x,r)) \le c_\mu r^m$ for all x and r, then for every sample and every $\sigma \le \tau\sqrt{d-m}/(6d)$, $W_1(\hat{\mu}_n * \mathcal{N}_\sigma, \mu_0) \ge c\, n^{-1/m}$, where $c = \frac{1}{20}(2c_\mu)^{-1/m}\sqrt{(d-m)/d}$. The isotropic kernel estimator thus does not beat the rate $n^{-1/m}$ of the empirical measure at any bandwidth in this range.*

We have no rate for mixing on $\mathcal{M}$ itself, and Remark H.1 states a conjecture. The proposition does not cover a kernel on a PCA subspace. Part (ii) applies within any affine subspace that contains $\mathcal{M}$, but a curved manifold need not lie in a low-dimensional one. The proof is in Section H.

5.3 WHEN k-NN SMOTE STOPS HELPING

The analysis also explains why SMOTE deteriorates in high dimension. With k-NN locality, our bound shows the gain of Section 3 only if $k > C_Q^d$, and this condition does not involve n (Corollary G.5). At the same time, locality itself erodes as d grows, since $h_\star/D \asymp (k/n)^{1/d}$ tends to one, so the nearest neighbour becomes an essentially arbitrary sample point and the bias is no longer small. The proved $C_Q = \sqrt{d}$ is a worst case, and if the effective constant is about 1.3 (Section G), the default $k = 5$ falls below C_Q^d from about $d = 6$. Beyond that point the bound shows mixing adding bias without reducing the variance. This agrees with the empirical findings of Blagus & Lusa (2013) and with the observation of Xia & Klusowski (2026) that resampling dominates SMOTE in that regime. On a k-NN graph the overlap stays within a constant factor of $1/n$ (Proposition G.1), so the failure comes from the loss of locality.

6 EXPERIMENTS

Setup. A first group of synthetic populations sweeps the arity from 2 to 12 on the flat torus and on a Gaussian, and tests the rate on a cosine torus. A second group compares mixing with the sample,

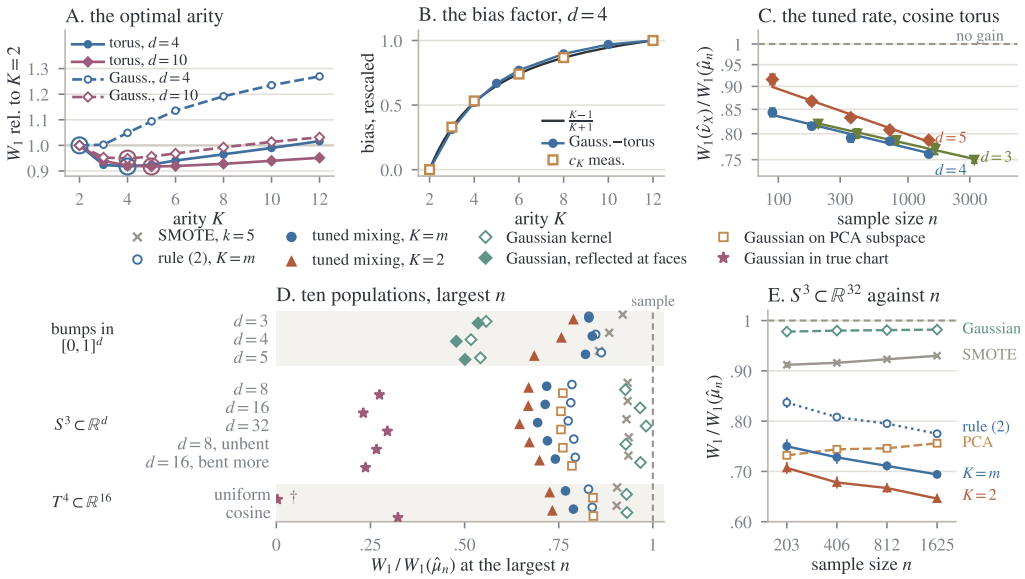

Figure 1: Top row: first group of experiments; bottom row: second. **A**: $W_1(\widehat{\nu}_X, \mu_0)$ relative to $K = 2$ on the flat torus ($\beta_1 \equiv 0$) and on a Gaussian, with 24 neighbours. Rings mark the best arity. **B**: the Gaussian-minus-torus gap of A at $d = 4$ and c_K measured directly, both rescaled to 0 at $K = 2$ and 1 at $K = 12$, against the $(K-1)/(K+1)$ of Proposition 4.1. **C**: the rule (2) over the empirical measure on the cosine torus, weighted fits. **D**: W_1 of each generator over that of the sample at the largest n, all ten populations. Here m is the intrinsic dimension, so $K = m$ means $K = d$ on the bumps, 3 on S^3 and 4 on T^4, and † marks the estimation floor. **E**: the same ratios against n on $S^3 \subset \mathbb{R}^{32}$. Floor removed in quadrature, bars one standard error.

SMOTE and kernel smoothers on ten populations in three families (Section I.2). The first family is four Gaussian bumps on a uniform floor in $[0,1]^d$ for $d = 3, 4, 5$, which meet Assumption 1. The second is a non-uniform density on the sphere S^3 bent into $\mathbb{R}^d$ for $d = 8, 16, 32$, and the third is a flat 4-torus in $\mathbb{R}^{16}$ with a uniform or a cosine density. We estimate W_1 between $N \geq 16384$ points of a generator and fresh points of μ_0 by a debiased Sinkhorn divergence, and we divide every score by that of the sample, which is the empirical measure of the same n points. The estimation floor, which is the score of a second fresh sample of μ_0, is 43–45% of the sample's score at the largest n and is removed in quadrature. Every comparison of levels below holds also with the floor left in (Section I.2). The scale of each generator, which is k or a bandwidth, is tuned against μ_0 on replicates that are not used for scoring. Since this tuning uses an oracle, we also report the untuned rule (2).

The gain over the sample (Figure 1C–E). Tuned pairs are 21–35% below the sample in every population and at every $n \geq 90$, and they are 17–27% below it at the largest n when the floor is left in, whereas SMOTE with its default $k = 5$ gains 6–15%. The rule (2) alone gains 14–23% at the largest n, which is more than $k = 5$ gains in every population but one, and in that one the two are level within a standard error. Its ratio to the sample falls with n in all ten populations, and in most of them it falls by more than two standard errors, as Theorem A predicts with m in place of d, although the fitted slopes are at most about half the predicted ones (Table 20). On the cosine torus the rule at $K = d$, which is used without the split, beats the sample by 8–25% at every n from 90 on, and its ratio to the sample falls with n in every d by 8 to 15 standard errors. The synthetic error itself falls as $n^{-0.39}$, $n^{-0.31}$ and $n^{-0.28}$ at $d = 3, 4, 5$, while Theorem A predicts $n^{-0.40}$, $n^{-0.33}$ and $n^{-0.29}$. The floor correction that gives the sample its exact slope lowers these exponents to 82–88% of the predicted ones (Table 11). The sign and the level of the gain are robust, but the slope of the ratio, which is a small difference of two exponents, is not settled at these n (Section I.1).

Arity and the bias factor (Figure 1A, B). On the flat torus $\mathbb{E}\widehat{\nu}_X = \mu_0$ at every K, so there is no mixing bias, and the error is smallest at $K = 4, 4, 5, 5$ for $d = 4, 6, 8, 10$, where it is 8–10% below that of pairs. The rise beyond comes from flat Dirichlet weights, which pull a K-fold mix towards the centroid of its tuple, and the penalty this causes crosses one between $K = 4$ and 5 in every dimension (Table 9). On a Gaussian the best arity is $K = 2, 3, 3, 4$ at the same d, and its gain over pairs is at most 5.2%. The gap between the two populations isolates the arity dependence of

the bias, and at $d = 4$ it follows the factor $(K-1)/(K+1)$ of Proposition 4.1 to within three points without any fitted parameter. The coefficient c_K measured directly meets it within 1.5%, and shrinking the neighbour pool moves the best arity as the bias factor predicts (Section I.1). In the second group, where k is tuned for each arity and $m \le 5$, pairs select a larger k than $K = m$ in all 39 cells (Table 23), and their error is 0.83–0.97 times that of $K = m$ at every n. Where two arities share $\bar{\kappa}$, only constants separate them. If the bias constant is that of Lemma E.5 with m in place of d and the fluctuation constants are equal, the balance behind (2) predicts a ratio of 0.938 for pairs against $K = 3$ at $m = 3$ and of 0.952 for $K = 3$ against $K = 4$ at $m = 4$. The measured means are 0.937 over 24 cells and 0.954 over 8 cells, although this check was made after the runs (Section I.2). The lead of pairs over $K = 4$ on the 4-torus grows with the bias, from 4.1% with the uniform density to 5.2% with the cosine one, and the difference in mean log ratio is $+0.011 \pm 0.003$. The proved exponents favour $K = 4$ there by a factor $n^{-1/21}$, which is invisible at $n \le 724$, and under Assumption 3 pairs lose only a logarithm. At these sample sizes, pairs are the practical choice.

Mixing against kernel smoothing (Figure 1D, E). On both manifolds, mixing beats the Euclidean Gaussian kernel at every n and codimension. The margin is 15–29% for $K = 3$ on S^3, 13–17% for $K = 4$ on T^4 and 15–34% for pairs. It widens with n in every population and with the codimension. At the largest n, the kernel's ratio to the sample rises from 0.93 to 0.97 and 0.98 as d goes from 8 to 32, while that of pairs stays at 0.65–0.67. By Proposition 5.1, the kernel keeps its first-order off-manifold bias small only by halving its bandwidth with each doubling of d, whereas mixing selects the same k at every d and bend (it does so in 39 of 40 selections, Table 23). A Gaussian kernel on the top principal subspace, which is given the dimension of the linear span, leads $K = m$ by 2–5% at the smallest n and is no better than pairs there. At the largest n, every arity leads it by 6–15%, and the gap widens with n in every population. A kernel that steps along the true chart of the manifold reaches 0.23–0.32 times the error of the sample at the largest n, and it reaches the estimation floor on the uniform torus. Mixing thus needs neither the dimension nor the subspace, but it is not the best smoother on a known manifold. In full dimension, a tuned Gaussian kernel, which has the rate of mixing on the torus (Lemma D.5), has the better constant. On the bumps the error of pairs is 1.25–1.59 times that of the kernel reflected at the faces, and the kernel also wins on the cosine torus (Section I.2). Its optimum is sharper. At $d = 4$, halving or doubling the bandwidth costs 6–41%, while halving or doubling k, which moves the locality by $2^{\pm 1/d}$, costs at most 8.5% (Table 24).

7 Limitations and Conclusion

A training set keeps the rate of the synthetic distribution only under the conditions of Corollary 3.1. The proofs split the sample, while the experiments run unsplit. Cross-fitting, which swaps $\mathcal{D}_A$ and $\mathcal{D}_B$ and averages the two synthetic distributions, keeps the rate by the joint convexity of W_1. Theorem A is proved for flat weights, and for $\mathrm{Dir}(\alpha)$ weights we have the bias at every K but a rate only for global pairs, and only towards their own population distribution (Remark D.3). The lower bound is proved on the torus, for a non-flat density and a bias of order h^2. The sharper exponents of Proposition E.10 rest on Assumption 3. Theorem A assumes a full-dimensional density bounded below, and on a manifold we have only Proposition 5.1 and the conjecture of Remark H.1. With the proved $C_Q = \sqrt{d}$, our bound beats the sample only for $n \gtrsim C_Q^{2d}$ (84), far above the measured crossover, since the effective C_Q is at most 1.3–1.5. The tuned W_2 rates are proved only on the torus, and our W_2 deviation bound for SMOTE is of order $n^{-1/4}$ (Remark F.2). We do not explain why a kernel does better in full-dimensional domains. All W_1 experiments are synthetic, the intrinsic dimension is at most 5 in the second group, and the scales are tuned against μ_0.

Convex augmentation is a smoother built from the data. When it is tuned, it beats the sample it is built from, and from $K = d - 1$ on it reaches $n^{-2/(d+2)}$, which no local convex scheme with a bias of order h^2 beats for a non-flat density on the torus. In practice, SMOTE's $k = 5$ should give way to a tuned k or the rule (2), with pairs at moderate n. Near a manifold, mixing beats an isotropic kernel unless the kernel knows the manifold. In full dimension, a tuned kernel does better.

AI USE STATEMENT

Because the full proofs are long, we used an AI assistant (Claude, Anthropic) to draft condensed proof sketches for the main text. The complete derivations appear unchanged in the Appendix. The authors checked every sketch against the full proof for correctness, rigor and clarity, and take full responsibility for all content. We also used it to edit the wording of the abstract and introduction.

REPRODUCIBILITY STATEMENT

Every result is stated with its assumptions and proved in the appendix. The reuse model and the W_1 fluctuation bound are in Section A, the smoothing refinement and the role of arity are in Sections B and C, and the mixing bias is in Section D. Section E proves Theorems A and B and the corollaries, Section F proves the W_2 bounds, Section G contains the k-NN analysis and Section H contains the manifold results. Tables 26 and 27 list every constant and exponent for which we give both a proved or predicted value and a measured one. Section I specifies every experiment (populations, generators, the divergence and its entropic parameter, the estimation floor and its correction, tuning grids, replicates and cloud sizes) and reports every run at every n with standard errors (Tables 14–24).

REFERENCES

I. S. Abramson. On bandwidth variation in kernel estimates—a square root law. *The Annals of Statistics*, 10(4):1217–1223, 1982.

L. Ambrosio, F. Stra, and D. Trevisan. A PDE approach to a 2-dimensional matching problem. *Probability Theory and Related Fields*, 173:433–477, 2019.

S. Bae, J.-W. Kim, W.-Y. Cho, H. Baek, S. Son, B. Lee, C. Ha, K. Tae, S. Kim, and S.-Y. Yun. Patch-mix contrastive learning with audio spectrogram transformer on respiratory sound classification. In *Proceedings of Interspeech 2023*, pp. 5436–5440, 2023.

R. Baena, L. Drumetz, and V. Gripon. Preventing manifold intrusion with locality: Local Mixup. *arXiv preprint arXiv:2201.04368*, 2022.

R. Blagus and L. Lusa. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14:106, 2013.

E. Boissard and T. Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'IHP Probabilités et Statistiques*, 50(2):539–563, 2014.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

L. Carratino, M. Cissé, R. Jenatton, and J.-P. Vert. On mixup regularization. *Journal of Machine Learning Research*, 23(325):1–31, 2022.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

H.-B. Chen and J. Niles-Weed. Asymptotics of smoothed Wasserstein distances. *Potential Analysis*, 56:571–595, 2022.

H. K. Choi, J. Choi, and H. J. Kim. TokenMixup: Efficient attention-guided token-level data augmentation for transformers. *Advances in Neural Information Processing Systems*, 35:14224–14235, 2022.

D. A. Dablain and N. V. Chawla. Towards understanding how data augmentation works with imbalanced data. *arXiv preprint arXiv:2304.05895*, 2023.

V. Divol. Measure estimation on manifolds: an optimal transport approach. *Probability Theory and Related Fields*, 183:581–647, 2022.

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

D. Elreedy and A. Atiya. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences*, 505:32–64, 2019.

D. Elreedy, A. F. Atiya, and F. Kamalov. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, 113:4903–4923, 2024.

H. Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

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

Z. Goldfeld, K. Greenewald, J. Weed, and Y. Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66 (7):4368–4391, 2020. arXiv:1905.13576.

M. Goldman, M. Huesmann, and F. Otto. Quantitative linearization results for the Monge–Ampère equation. *Communications on Pure and Applied Mathematics*, 74(12):2483–2560, 2021.

K. Greenewald, A. Gu, M. Yurochkin, J. Solomon, and E. Chien. k-Mixup regularization for deep learning via optimal transport. *Transactions on Machine Learning Research*, 2023.

L. Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

H. Guo, Y. Mao, and R. Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3714–3722, 2019.

O. Kachan, A. Savchenko, and G. Gusev. Simplicial SMOTE: Oversampling solution to the imbalanced learning problem. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025.

F. Kamalov, H. Sulieman, and W. Pedrycz. Theoretical convergence of SMOTE-generated samples. *arXiv preprint arXiv:2601.01927*, 2026.

W. H. Kang, J. Alam, and A. Fathan. L-mix: A latent-level instance mixup regularization for robust self-supervised speaker representation learning. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1263–1272, 2022.

M. Kimura. Why mixup improves the model performance. In *International Conference on Artificial Neural Networks*, pp. 275–286. Springer, 2021.

L. Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

M. Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.

T. Liang. How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021.

P. Mangla, V. Singh, S. Havaldar, and V. Balasubramanian. On the benefits of defining vicinal distributions in latent space. *Pattern Recognition Letters*, 152:382–390, 2021.

J. Niles-Weed and Q. Berthet. Minimax estimation of smooth densities in Wasserstein distance. *The Annals of Statistics*, 50(3):1519–1540, 2022.

P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1–3):419–441, 2008.

K. Oko, S. Akiyama, and T. Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pp. 26517–26582. PMLR, 2023.

R. Peyre. Comparison between W_2 distance and $\dot{H}^{-1}$ norm, and localization of Wasserstein distance. *ESAIM: Control, Optimisation and Calculus of Variations*, 24(4):1489–1501, 2018.

F. Pinto, H. Yang, S. N. Lim, P. Torr, and P. Dokania. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. *Advances in Neural Information Processing Systems*, 35:14608–14622, 2022.

A. Sakho, E. Malherbe, and E. Scornet. Do we need rebalancing strategies? A theoretical and empirical study around SMOTE and its variants. *arXiv preprint arXiv:2402.03819*, 2024.

T. Shimada, S. Yamaguchi, K. Hayashi, and S. Kobayashi. Data interpolating prediction: Alternative interpretation of mixup. *arXiv preprint arXiv:1906.08412*, 2019.

B. W. Silverman and G. A. Young. The bootstrap: To smooth or not to smooth? *Biometrika*, 74(3): 469–479, 1987.

S. Stocksieker, D. Pommeret, and A. Charpentier. Generalized oversampling for learning from imbalanced datasets and associated theory. *arXiv preprint arXiv:2308.02966*, 2023.

S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pp. 6438–6447. PMLR, 2019.

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.

J. Weed and Q. Berthet. Estimation of smooth densities in Wasserstein distance. In *Conference on Learning Theory (COLT)*, volume 99 of *PMLR*, pp. 3118–3119, 2019.

E. Xia and J. M. Klusowski. Classification imbalance as transfer learning. *arXiv preprint arXiv:2601.10630*, 2026.

X. Xie, Y. Li, W. Chen, K. Ouyang, Z. Xie, and H. T. Zheng. Global mixup: Eliminating ambiguity with clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37 (11), pp. 13798–13806, 2023.

C. Zhang, M.-H. Hsieh, and D. Tao. Generalization bounds for vicinal risk minimization principle. *arXiv preprint arXiv:1811.04351*, 2018a.

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018b.

L. Zhang, Z. Deng, K. Kawaguchi, A. Ghorbani, and J. Zou. How does mixup help with robustness and generalization? In *International Conference on Learning Representations*, 2021.

L. Zhang, Z. Deng, K. Kawaguchi, and J. Zou. When and how mixup improves calibration. In *International Conference on Machine Learning*, pp. 26135–26160. PMLR, 2022.

APPENDIX

The appendix contains complete proofs. Section A sets out the general reuse model, its overlap factor and its concentration, and proves Theorem A.7, Corollary A.10 and Proposition A.11; Section B states and proves the smoothing refinement, Theorem B.3, with Proposition B.5 and Proposition B.4; Section C proves Assumption 2 for pairs and neighbourhood graphs, and the torus bound of Theorem F.1(i) for (G1); Section D bounds the mixing bias; Section E proves Theorem A and Theorem A′, the lower bound and the corollaries of Section 3, and Theorem F.1 for (G2); Section F states the W_2 counterpart of Theorem A′ and proves it for (G1); Section G treats SMOTE in practice, from its overlap and the splitting repair to tuning and the measured constants; Section H treats data on a manifold; Section I gives the experimental details; and Section J collects a comparison of the methods and a table of proved against measured constants.

A THE GENERAL REUSE MODEL

A.1 THE MODEL AND THE OVERLAP FACTOR

Let $X_1, \dots, X_n$ be independent random vectors in $\mathbb{R}^d$ with $X_i \sim \mu_i$, and let $\mu := \bigotimes_{i=1}^n \mu_i$ be their joint distribution on $(\mathbb{R}^d)^n$, equipped with the concatenated Euclidean norm $\|X\| := (\sum_{i=1}^n \|X_i\|_2^2)^{1/2}$. In the i.i.d. statements below all μ_i equal a common distribution μ_0 with $\operatorname{diam}(\operatorname{supp}\mu_0) \le D$.

Definition A.1 (Reuse generator)**.** Fix $K \ge 1$ ($K = 1$ is the bootstrap) and a distribution π on index tuples $\{1, \dots, n\}^K$, and draw $\alpha^1, \dots, \alpha^{n_\pi} \overset{\text{i.i.d.}}{\sim} \pi$ *with replacement*, independently of the data, so that n_π may exceed the number of distinct tuples available; A denotes the multiset of the n_π draws. For each draw take $u^\alpha = (u_1^\alpha, \dots, u_K^\alpha) \sim \mathrm{Dir}(1, \dots, 1)$, independently across draws and of the data, and set

$$Y_\alpha(X, u^\alpha) \;:=\; \sum_{j=1}^K u_j^\alpha X_{\alpha_j}. \tag{13}$$

The *conditional synthetic distribution*, the *reuse-weighted empirical measure* and the *population synthetic distribution* are

$$\widehat{\nu}_X := \frac{1}{n_\pi} \sum_{t=1}^{n_\pi} \mathcal{L}\big(Y_{\alpha^t}(X, u^{\alpha^t}) \mid X\big), \qquad \mu_w := \sum_{i=1}^n w_i \delta_{X_i}, \qquad \bar{\nu} := \mathbb{E}\,\widehat{\nu}_X. \tag{14}$$

Definition A.2 (Multiplicity, weight, overlap factor)**.** For each i let $r_i := \#\{(t, j) : t \le n_\pi,\ j \le K,\ \alpha_j^t = i\}$ be its *multiplicity*, let $w_i := r_i/(n_\pi K)$ be its *weight*, so that $w_i \ge 0$ and $\sum_i w_i = 1$, and let

$$S \;:=\; \sum_{i=1}^n w_i^2 = \sum_{i=1}^n \Big(\frac{r_i}{n_\pi K}\Big)^2 \tag{15}$$

be the *overlap factor*. Sampling with replacement makes w random: writing p_i for the expected share of one draw and $S_\pi := \sum_i p_i^2$ for the *population overlap* of π, one has $\mathbb{E}S \le S_\pi + 1/n_\pi$, so drawing more points does not inflate reuse but settles it at S_π (Section G). When log-Sobolev constants α_i are available we also use the proxy $V_1 := \sum_i \alpha_i w_i^2 \le \alpha S$ if $\alpha_i \le \alpha$.

Every bound below is stated for a realised w, hence holds conditionally on the draw; $\bar{\nu}$ does not depend on it, and ψ_d is concave on $(0, 1]$, ψ_2 included, so averaging replaces S by $\mathbb{E}S$. A data-built π is the subject of Theorem A.

We write $\mathcal{S}_K(\rho) := \mathcal{L}(\sum_{j=1}^K U_j Z_j)$ for the *population synthetic operator*, where $Z_1, \dots, Z_K \overset{\text{i.i.d.}}{\sim} \rho$ and $U \sim \mathrm{Dir}(1, \dots, 1)$ is independent of them, and set $\nu_{K,w} := \mathcal{S}_K(\mu_w)$. If the indices inside every tuple are pairwise distinct and the data are i.i.d., then $\bar{\nu} = \mathcal{S}_K(\mu_0)$.

Example A.3 (mixup and SMOTE)**.** For mixup, π draws a pair uniformly, so $S_\pi = 1/n$, the minimum. For SMOTE, π completes a point with neighbours in the ε-graph, and *which* point is drawn uniformly decides the answer: a uniform base keeps S_π near $1/n$ (Proposition G.1), whereas a uniform K-clique weights each point by the number of cliques containing it, hence by the local density, and inflates S_π.

Remark A.4 (Lower bound on overlap). By Cauchy–Schwarz, $S \ge \frac{1}{n}\big(\sum_i w_i\big)^2 = \frac{1}{n}$ for every multiset A of draws and every K, with equality if and only if $w_i = 1/n$ for all i, i.e. reuse is perfectly balanced.

Every rate below increases with S and at $S = 1/n$ reproduces the empirical-measure rate, so $n_{\text{eff}} := 1/S \le n$ is the effective sample size of the synthetic pool. Hence:

Remark A.5 (No free information). Synthetic convex combinations cannot improve the intrinsic n-dependent *fluctuation* scale, however many are generated, and imbalanced reuse is strictly harmful; their value is in the smoothing, which buys a better *total* error by paying bias, not by extracting information (Theorem A).

All our Wasserstein statements compare against the population distribution μ_0 through the split (1), which holds for every $p \ge 1$. The second term is deterministic and carries *no* dependence on n, n_π or

S; all randomness and all overlap dependence live in the first. In W_1 the fluctuation is measured by the transportation profile

$$\psi_d(S) := S^{1/2}\ (d=1), \qquad S^{1/2}\big(1+\log(1/S)\big)\ (d=2), \qquad S^{1/d}\ (d\geq 3), \tag{16}$$

with the explicit constants $C_1 = \frac{1}{2}$, $C_2 = 2\sqrt{2}$ and $C_d = 2\sqrt{d}\,(2r-1)/(r-1)$, $r := 2^{d/2-1}$, for $d \geq 3$; applied to $\widehat{\nu}_X$ they are multiplied by $\sqrt{b_K} \leq K$ (Section A.4 Step 7). The W_2 counterpart of Theorem A′, for the same generators, is Theorem F.1 (Section F).

A.2 Concentration, the exact split, and the W_1 bound

Concentration under reuse. For $f : \mathbb{R}^d \to \mathbb{R}$ 1-Lipschitz put $g_\alpha(X) := \mathbb{E}_{u^\alpha}[f(Y_\alpha(X, u^\alpha)) \mid X]$ and $T_f(X) := \frac{1}{n_\pi}\sum_{\alpha\in A}(g_\alpha(X) - \mathbb{E}g_\alpha(X))$, the centred synthetic functional. The whole dependence structure of the generator is compressed by the following deterministic lemma, proved in Section A.3.

Lemma A.6 (Lipschitz profile of the generator)**.** *For every 1-Lipschitz f, the map T_f is coordinatewise Lipschitz with constants $c_i = w_i$, i.e. $|T_f(X) - T_f(X')| \leq \sum_{i=1}^n w_i \|X_i - X_i'\|_2$, and consequently $\mathrm{Lip}(T_f) \leq \sqrt{S}$ with respect to $\|\cdot\|$.*

On a set of diameter D, changing X_i moves T_f by at most $w_i D$, so McDiarmid's inequality makes T_f sub-Gaussian with proxy $V := D^2 S/4$, and every deviation bound below holds with this V and no further hypothesis. A log-Sobolev inequality for the coordinates gives the proxy $V_1 = \sum_i \alpha_i w_i^2$ instead and needs no support bound (Proposition A.8); it serves the Gaussian population of Section 6 and the W_2 deviation of Section F (Remark A.9). Either way n_π enters only through w: more combinations from the same pool do not sharpen concentration.

The exact fluctuation–bias split. Throughout the rest of this subsection, $X_1, \ldots, X_n$ are i.i.d. μ_0 with $\mathrm{diam}(\mathrm{supp}\,\mu_0) \leq D$. The dyadic estimates follow the empirical-measure proofs (Fournier & Guillin, 2015; Weed & Bach, 2019), except that a cell mass has variance $\leq Sp_Q$ rather than $p_Q(1-p_Q)/n$: replacing $1/n$ by S is how overlap enters every rate. Since $\bar{\nu} = \mathbb{E}\widehat{\nu}_X$ is deterministic, Kantorovich–Rubinstein duality gives

$$W_1(\widehat{\nu}_X, \bar{\nu}) \;=\; \sup_{\mathrm{Lip}(f)\leq 1} T_f(X), \tag{17}$$

so the fluctuation term is *exactly* the supremum of the centred process of Lemma A.6, which removes the interpolation-bias term $\mathbb{E}W_1(\widehat{\nu}_X, \mu_w)$ a μ_w-based route would carry.

Theorem A.7 (W_1 bound, fluctuation–bias form)**.** *Assume the indices inside every tuple are pairwise distinct, so that $\bar{\nu} = \mathcal{S}_K(\mu_0)$, and set $\beta_1 := W_1(\bar{\nu}, \mu_0)$. Then*

$$\mathbb{E}\, W_1(\widehat{\nu}_X, \bar{\nu}) \;\leq\; C_d D\, \psi_d(S), \tag{18}$$

the map $X \mapsto W_1(\widehat{\nu}_X, \bar{\nu})$ is coordinatewise Lipschitz with constants $c_i = w_i$, and for every $t \geq 0$, $\mathbb{P}(W_1(\widehat{\nu}_X, \bar{\nu}) \geq \mathbb{E}W_1(\widehat{\nu}_X, \bar{\nu}) + t) \leq \exp(-t^2/2V)$. Consequently, with probability at least $1-\delta$,

$$W_1(\widehat{\nu}_X, \mu_0) \;\leq\; \beta_1 \;+\; C_d D\, \psi_d(S) \;+\; \sqrt{2V\log(1/\delta)}. \tag{19}$$

A.3 Concentration under reuse: the Lipschitz profile and the sub-Gaussian bound

Proposition A.8 (Concentration under reuse)**.** *Assume each μ_i satisfies a logarithmic Sobolev inequality with constant $\alpha_i > 0$. Then for every $\lambda \in \mathbb{R}$, $\mathbb{E}\exp(\lambda(T_f(X) - \mathbb{E}T_f(X))) \leq \exp(\lambda^2 V_1/2)$ with $V_1 = \sum_{i=1}^n \alpha_i w_i^2$, and therefore, for every $t \geq 0$,*

$$\mathbb{P}\big(T_f(X) \geq \mathbb{E}T_f(X) + t\big) \;\leq\; \exp\Big(-\frac{t^2}{2V_1}\Big) \;\leq\; \exp\Big(-\frac{t^2}{2\alpha S}\Big) \quad \textit{whenever } \alpha_i \leq \alpha\ \forall i. \tag{20}$$

Remark A.9 (When the log-Sobolev inequality is needed). On a set of diameter D it is not: by Lemma A.6 a change of X_i moves T_f by at most $c_i = w_i D$, and so it moves every map that inherits that profile (Section A.4 Step 8); McDiarmid's inequality then gives the sub-Gaussian proxy

$\sum_i c_i^2/4 = D^2 S/4$ with no hypothesis on the μ_i. For μ_0 uniform on a convex body the two proxies are of the same order, since then $\alpha \asymp D^2$. Proposition A.8 matters for unbounded supports, the Gaussian population of Section I with $\alpha_i = \sigma^2$, and for the W_2 deviation of Theorem F.1(ii) in (G1): there bounded differences give constants $\sqrt{w_i} D$, whose squares sum to D^2 and do not decay (Remark F.7), and the transport-weighted gradient of Lemma F.6 needs the Herbst argument.

The proof of Proposition A.8 has two ingredients: T_f is $\sqrt{S}$-Lipschitz as a function of $X = (X_1, \dots, X_n) \in (\mathbb{R}^d)^n$, and the log-Sobolev inequality upgrades Lipschitzness to sub-Gaussian concentration through the Herbst argument.

Step 1: coordinatewise sensitivity. Fix $i \in \{1, \dots, n\}$ and consider X, X' that differ only in coordinate i: $X'_j = X_j$ for $j \neq i$ and $X'_i = X_i + h$. Fix $\alpha \in A$, let $J_{\alpha,i} := \{j \leq K : \alpha_j = i\}$ and $c_{\alpha,i} := |J_{\alpha,i}|$. Then

$$Y_\alpha(X) = \sum_{j \notin J_{\alpha,i}} u_j^\alpha X_{\alpha_j} + \sum_{j \in J_{\alpha,i}} u_j^\alpha X_i, \qquad Y_\alpha(X') = Y_\alpha(X) + \Big(\sum_{j \in J_{\alpha,i}} u_j^\alpha \Big) h.$$

Since f is 1-Lipschitz, $|f(Y_\alpha(X')) - f(Y_\alpha(X))| \leq (\sum_{j \in J_{\alpha,i}} u_j^\alpha) \|h\|_2$. Taking $\mathbb{E}_{u^\alpha}[\,\cdot \mid X]$ and using $\mathbb{E} u_j^\alpha = 1/K$ for $u^\alpha \sim \mathrm{Dir}(1, \dots, 1)$,

$$|g_\alpha(X') - g_\alpha(X)| \;\leq\; \mathbb{E}\Big[\sum_{j \in J_{\alpha,i}} u_j^\alpha \Big] \|h\|_2 \;=\; \frac{c_{\alpha,i}}{K} \|h\|_2.$$

Subtracting $\mathbb{E} g_\alpha(X)$ does not change Lipschitz constants, so averaging over $\alpha \in A$ and using $\sum_{\alpha \in A} c_{\alpha,i} = r_i$,

$$|T_f(X') - T_f(X)| \;\leq\; \frac{1}{n_\pi} \sum_{\alpha \in A} \frac{c_{\alpha,i}}{K} \|h\|_2 \;=\; \frac{r_i}{n_\pi K} \|h\|_2 \;=\; w_i \|h\|_2. \tag{21}$$

Step 2: global Lipschitz constant. Let $X, X' \in (\mathbb{R}^d)^n$ be arbitrary and $\Delta_i := X_i - X'_i$. Replacing coordinates one at a time and applying (21) at each step,

$$|T_f(X) - T_f(X')| \;\leq\; \sum_{i=1}^n w_i \|\Delta_i\|_2 \;\leq\; \Big(\sum_{i=1}^n w_i^2 \Big)^{1/2} \Big(\sum_{i=1}^n \|\Delta_i\|_2^2 \Big)^{1/2} \;=\; \sqrt{S}\, \|X - X'\|,$$

by Cauchy–Schwarz. This proves Lemma A.6; equivalently, by Rademacher's theorem T_f is a.e. differentiable with $\|\nabla T_f(X)\|_2 \leq \sqrt{S}$.

Step 3: log-Sobolev for the product measure. Each μ_i satisfies $\mathrm{Ent}_{\mu_i}(e^h) \leq \frac{\alpha_i}{2} \int \|\nabla h\|_2^2 e^h \, d\mu_i$ for all smooth h (Gross, 1975). By tensorisation (Ledoux, 2001), $\mu = \bigotimes_i \mu_i$ satisfies an LSI with constant $\alpha := \max_i \alpha_i$, and, handling the product constant coordinatewise, for a function with coordinatewise Lipschitz constants c_i the effective constant is $\sum_i \alpha_i c_i^2$.

Step 4: Herbst argument. This is the entropy method of Boucheron et al. (2013). Apply the LSI to $H = \lambda T_f$. Using $\nabla(\lambda T_f) = \lambda \nabla T_f$ and the coordinatewise bound of Step 1,

$$\mathrm{Ent}_\mu(e^{\lambda T_f}) \;\leq\; \frac{\lambda^2}{2} \Big(\sum_i \alpha_i w_i^2 \Big) \int e^{\lambda T_f} d\mu \;=\; \frac{\lambda^2 V_1}{2} \int e^{\lambda T_f} d\mu.$$

Let $\psi(\lambda) := \log \mathbb{E}[e^{\lambda T_f(X)}]$. The standard identity $\mathrm{Ent}(e^{\lambda T_f}) = \mathbb{E}[e^{\lambda T_f}](\lambda \psi'(\lambda) - \psi(\lambda))$ and division by $\mathbb{E}[e^{\lambda T_f}]$ give $\lambda \psi'(\lambda) - \psi(\lambda) \leq \lambda^2 V_1/2$. With $\phi(\lambda) := \psi(\lambda)/\lambda$ this reads $\phi'(\lambda) = \frac{\lambda \psi'(\lambda) - \psi(\lambda)}{\lambda^2} \leq V_1/2$. Integrating from 0 to λ and using $\lim_{\lambda \to 0} \phi(\lambda) = \mathbb{E} T_f(X)$,

$$\psi(\lambda) \;\leq\; \lambda \, \mathbb{E} T_f(X) + \frac{\lambda^2 V_1}{2}, \qquad \text{i.e.} \qquad \mathbb{E} \exp\big(\lambda (T_f - \mathbb{E} T_f)\big) \leq \exp\Big(\frac{\lambda^2 V_1}{2} \Big).$$

Step 5: Chernoff optimisation. For $t \geq 0$ and $\lambda > 0$, Markov's inequality gives $\mathbb{P}(T_f - \mathbb{E} T_f \geq t) \leq \exp(-\lambda t + \lambda^2 V_1/2)$, minimised at $\lambda^\star = t/V_1$, yielding $\exp(-t^2/(2V_1))$. A symmetric bound for the lower tail follows by applying the same argument to $-T_f$. If $\alpha_i \leq \alpha$ then $V_1 \leq \alpha \sum_i w_i^2 = \alpha S$. For $X_i \sim \mathcal{N}(0, I_d)$ the product distribution on $\mathbb{R}^{nd}$ is standard Gaussian and satisfies an LSI with constant 1, so $\alpha = 1$ and $V_1 = S$. □

A.4 Proof of the W_1 fluctuation–bias bound

Step 1: exact dual identity. Since $\bar{\nu} = \mathbb{E}\widehat{\nu}_X$ is deterministic, Kantorovich–Rubinstein duality gives $W_1(\widehat{\nu}_X, \bar{\nu}) = \sup_{\mathrm{Lip}(f)\le 1}[\widehat{\nu}_X(f) - \bar{\nu}(f)]$. Writing $\widehat{\nu}_X(f) = \frac{1}{n_\pi}\sum_\alpha g_\alpha(X)$ and $\bar{\nu}(f) = \frac{1}{n_\pi}\sum_\alpha \mathbb{E}g_\alpha(X)$ we get, for each fixed f, $\widehat{\nu}_X(f) - \bar{\nu}(f) = T_f(X)$, hence the exact identity (17). No interpolation-bias term arises: routing through $\bar{\nu}$ replaces the additive $\eta_{\mathrm{int},1} = \mathbb{E}W_1(\widehat{\nu}_X, \mu_w)$ of a μ_w-based argument by a pure fluctuation supremum.

Step 2: the one-dimensional case. Assume $d = 1$ and, after translation, $\mathrm{supp}\,\mu_0 \subset [0, D]$. Let F be the c.d.f. of μ_0 and $F_w(t) := \sum_i w_i \mathbf{1}_{\{X_i \le t\}}$; then $W_1(\mu_w, \mu_0) = \int_0^D |F_w(t) - F(t)|\, dt$. For fixed t, $\mathbb{E}F_w(t) = F(t)$ and, by independence, $\mathrm{Var}(F_w(t)) = \sum_i w_i^2 F(t)(1 - F(t)) \le \frac{1}{4}\sum_i w_i^2 = S/4$, so Jensen gives $\mathbb{E}|F_w(t) - F(t)| \le \sqrt{S}/2$ and hence $\mathbb{E}W_1(\mu_w, \mu_0) \le D\sqrt{S}/2$, i.e. $C_1 = 1/2$.

Step 3: dyadic transportation estimate for $d \ge 2$. After translation assume $\mathrm{supp}\,\mu_0 \subset Q_0$, a cube of side D. For $\ell \ge 0$ let $\mathcal{Q}_\ell$ be the dyadic partition of Q_0 into $2^{d\ell}$ cubes of side $D2^{-\ell}$, and for $Q \in \mathcal{Q}_\ell$ set $\Delta_Q := \mu_w(Q) - \mu_0(Q)$ and $A_\ell := \sum_{Q\in\mathcal{Q}_\ell} |\Delta_Q|$. The standard dyadic estimate states that for every integer $L \ge 1$,

$$W_1(\mu_w, \mu_0) \;\le\; \sqrt{d}D2^{-L} + \sqrt{d}D\sum_{\ell=1}^{L} 2^{-(\ell-1)} A_\ell. \tag{22}$$

One first matches the two measures inside the cells of $\mathcal{Q}_L$, at a within-cell cost of at most $\sqrt{d}D2^{-L}$; the remaining unmatched mass is then transported recursively between children of dyadic parents, where at level ℓ the parent diameter is at most $\sqrt{d}D2^{-(\ell-1)}$ and the redistributed mass is bounded by A_ℓ.

Step 4: expected mass discrepancy. Write $p_Q := \mu_0(Q)$. Since $\mu_w(Q) = \sum_i w_i \mathbf{1}_{\{X_i \in Q\}}$ we have $\mathbb{E}\mu_w(Q) = p_Q$ and, by independence, $\mathrm{Var}(\mu_w(Q)) = \sum_i w_i^2 p_Q(1 - p_Q) \le S p_Q$. Jensen's inequality gives $\mathbb{E}|\Delta_Q| \le \sqrt{S p_Q}$, and summing over $Q \in \mathcal{Q}_\ell$ with Cauchy–Schwarz,

$$\mathbb{E}A_\ell \;\le\; \sqrt{S}\sum_{Q\in\mathcal{Q}_\ell}\sqrt{p_Q} \;\le\; \sqrt{S}\,(\#\mathcal{Q}_\ell)^{1/2}\Big(\sum_Q p_Q\Big)^{1/2} \;=\; \sqrt{S}\,2^{d\ell/2}. \tag{23}$$

Substituting into (22), $\mathbb{E}W_1(\mu_w, \mu_0) \le \sqrt{d}D2^{-L} + 2\sqrt{d}D\sqrt{S}\sum_{\ell=1}^{L} 2^{\ell(d/2-1)}$.

Step 5: optimisation for $d = 2$. Here $\sum_{\ell\le L} 2^{\ell(d/2-1)} = L$. For $0 < S < 1$ choose $L = \lfloor \frac{1}{2}\log_2 \frac{1}{S} \rfloor$ (and $L = 1$ if $S = 1$), so that $2^{-L} \le 2\sqrt{S}$. Then $\mathbb{E}W_1(\mu_w, \mu_0) \le 2\sqrt{2}D\sqrt{S} + 2\sqrt{2}D\sqrt{S}L \le 2\sqrt{2}D\sqrt{S}(1 + \log\frac{1}{S})$, i.e. $C_2 = 2\sqrt{2}$.

Step 6: optimisation for $d \ge 3$. Let $r := 2^{d/2-1} > 1$ and choose $L = \lfloor \frac{1}{d}\log_2 \frac{1}{S} \rfloor$ (and $L = 1$ if $S = 1$), so that $2^{-L} \le 2S^{1/d}$. Then

$$\sqrt{S}\sum_{\ell=1}^{L} 2^{\ell(d/2-1)} = \sqrt{S}\sum_{\ell=1}^{L} r^\ell \le \sqrt{S}\,\frac{r^{L+1}}{r-1} \le \frac{r}{r-1}S^{1/d},$$

whence $\mathbb{E}W_1(\mu_w, \mu_0) \le 2\sqrt{d}D\,(1 + \frac{r}{r-1})S^{1/d}$, i.e. $C_d = 2\sqrt{d}\,(2r-1)/(r-1)$ with $r = 2^{d/2-1}$.

Step 7: from μ_w to $\widehat{\nu}_X$. The same dyadic argument applies verbatim with μ_0 replaced by the deterministic measure $\bar{\nu}$ and μ_w replaced by the smoothed synthetic measure $\widehat{\nu}_X$: the argument only uses $\mathbb{E}\Delta_Q = 0$ and a variance bound of the form $\mathrm{Var}(\widehat{\nu}_X(Q)) \le b_K S\,\bar{\nu}(Q)$. To obtain it, put $Z_\alpha := \rho_{\alpha,X}(Q) \in [0, 1]$. With i.i.d. data and pairwise distinct indices every component has the same mean, $\mathbb{E}Z_\alpha = \bar{\nu}(Q)$, so $|\mathrm{Cov}(Z_\alpha, Z_\beta)| \le (\mathbb{E}Z_\alpha \mathbb{E}Z_\beta)^{1/2} = \bar{\nu}(Q)$. Two components are independent unless their tuples share an index, and $\sum_\alpha \#\{\beta : \beta \sim \alpha\} \le \sum_\alpha \sum_{j\le K} r_{\alpha_j} = \sum_i r_i^2 = (n_\pi K)^2 S$, so

$$\mathrm{Var}(\widehat{\nu}_X(Q)) \;\le\; \frac{\bar{\nu}(Q)}{{n_\pi}^2}\sum_\alpha \#\{\beta : \beta \sim \alpha\} \;\le\; K^2 S\,\bar{\nu}(Q),$$

i.e. $b_K \le K^2$; for the unsmoothed empirical measure Step 4 gives $b_1 = 1$ exactly. Hence (18) holds with C_d multiplied by $\sqrt{b_K} \le K$. The constants C_d are therefore explicit for μ_w and explicit up to this factor for $\widehat{\nu}_X$; numerically $\sqrt{b_K}$ stays well below K but does grow with it under locality: in $d = 1$ at $K = 2, 3, 5$ it is $0.85, 0.78, 0.84$ for global mixing and $1.47, 1.86, 2.60$ for $h_\star$-local mixing with cells of side $2h_\star$.

Step 8: Lipschitz property and concentration. Put $H(X) := W_1(\widehat{\nu}_X, \bar{\nu})$. Since $\bar{\nu}$ does not depend on X, the triangle inequality and duality give $|H(X) - H(X')| \le W_1(\widehat{\nu}_X, \widehat{\nu}_{X'}) \le \frac{1}{n_\pi}\sum_\alpha \mathbb{E}_{u^\alpha}\|Y_\alpha(X, u^\alpha) - Y_\alpha(X', u^\alpha)\|_2$. Using $Y_\alpha(X, u^\alpha) - Y_\alpha(X', u^\alpha) = \sum_j u_j^\alpha (X_{\alpha_j} - X'_{\alpha_j})$, the triangle inequality and $\mathbb{E}u_j^\alpha = 1/K$,

$$\frac{1}{n_\pi}\sum_{\alpha\in A}\mathbb{E}_{u^\alpha}\Big\|\sum_j u_j^\alpha (X_{\alpha_j} - X'_{\alpha_j})\Big\|_2 \le \frac{1}{n_\pi}\sum_{\alpha\in A}\frac{1}{K}\sum_{j=1}^{K}\|X_{\alpha_j} - X'_{\alpha_j}\|_2$$
$$= \sum_{i=1}^{n}\frac{r_i}{n_\pi K}\|X_i - X'_i\|_2 = \sum_{i=1}^{n} w_i\|X_i - X'_i\|_2.$$

Hence H is coordinatewise Lipschitz with $c_i = w_i$, so on a set of diameter D it has bounded differences $w_i D$ and McDiarmid's inequality gives $\mathbb{E}\exp(\lambda(H - \mathbb{E}H)) \le \exp(\lambda^2 V/2)$ and $\mathbb{P}(H \ge \mathbb{E}H + t) \le \exp(-t^2/(2V))$ with $V = D^2 S/4$ (Remark A.9); under log-Sobolev inequalities Steps 3–5 of Section A.3 give the same with $V = V_1$. Equivalently, with probability at least $1 - \delta$, $H \le \mathbb{E}H + \sqrt{2V\log(1/\delta)}$. Combining with (1) and (18) yields (19); if $\alpha_i \le \alpha$ then $V_1 \le \alpha S$. □

Corollary A.10 (Synthetic-population bounds, bias-free)**.** *If the object of interest is the synthetic distribution itself rather than μ_0, the mixing bias disappears and only the overlap term remains: with $\nu_{K,w} = \mathcal{S}_K(\mu_w)$ and $\nu_{K,0} = \mathcal{S}_K(\mu_0)$, $\mathbb{E}W_1(\nu_{K,w}, \nu_{K,0}) \le C_d D\psi_d(S)$ and $\mathbb{E}W_2(\nu_{K,w}, \nu_{K,0}) \le \widetilde{C}_d D\varphi_d(S)$; with probability at least $1 - \delta$ the same bounds hold with the additional terms $\sqrt{2V_1\log(1/\delta)}$ and $\sqrt{2V_2\log(1/\delta)}$ respectively (φ_d, $\widetilde{C}_d$ and V_2 as in Section F).*

Proof of Corollary A.10. Apply Steps 2–6 with μ_w and μ_0 replaced by $\nu_{K,w} = \mathcal{S}_K(\mu_w)$ and $\nu_{K,0} = \mathcal{S}_K(\mu_0)$; since $\mathcal{S}_K$ is a 1-Lipschitz map for W_p in the sense that $W_p(\mathcal{S}_K(\rho), \mathcal{S}_K(\rho')) \le W_p(\rho, \rho')$ (couple the Z_j's optimally and share the same U), the expectation bounds are inherited. The deviation bounds follow from Step 8 and, for W_2, from Section F Step 7. □

Proposition A.11 (Excess risk of training on synthetic data)**.** *Let $\{\ell(\cdot, \theta)\}_{\theta\in\Theta}$ be L-Lipschitz in x uniformly in θ and let $\hat{\theta}$ minimise the synthetic risk $R_{\widehat{\nu}_X}$. Then $R_{\mu_0}(\hat{\theta}) - \inf_{\theta\in\Theta} R_{\mu_0}(\theta) \le 2L\, W_1(\widehat{\nu}_X, \mu_0)$, so (19), (29) and Theorem A′ bound the excess risk by the same terms, scaled by $2L$.*

Proof of Proposition A.11. For any θ, $|R_{\widehat{\nu}_X}(\theta) - R_{\mu_0}(\theta)| \le L\, W_1(\widehat{\nu}_X, \mu_0)$ by Kantorovich–Rubinstein duality, since $\ell(\cdot, \theta)/L$ is 1-Lipschitz. For θ' with $R_{\mu_0}(\theta') \le \inf R_{\mu_0} + \eta$, $R_{\mu_0}(\hat{\theta}) \le R_{\widehat{\nu}_X}(\hat{\theta}) + LW_1 \le R_{\widehat{\nu}_X}(\theta') + LW_1 \le R_{\mu_0}(\theta') + 2LW_1$; let $\eta \downarrow 0$. □

Remark A.12 (Unbounded support). Theorem A.7 is stated for $\mathrm{diam}(\mathrm{supp}\,\mu_0) \le D$. The expectation bound extends to distributions with a finite moment $M_q := \mathbb{E}\|X\|^q$ by the shell decomposition of Fournier & Guillin (2015), whose proof uses the sample only through $\mathbb{E}|\mu_N(B) - \mu(B)| \le \min\{2\mu(B), (\mu(B)/N)^{1/2}\}$ for Borel B. For μ_w this holds with $1/N$ replaced by S (Step 4), and for $\widehat{\nu}_X$ with $b_K S$ (Step 7), while $\bar{\nu}$ inherits M_q from μ_0 by convexity. Hence for $d \ge 3$ and $q > d/(d-1)$, $\mathbb{E}W_1(\widehat{\nu}_X, \bar{\nu}) \le C_{d,q} M_q^{1/q} (b_K S)^{1/d}$, and a Gaussian, with all moments finite, has the baseline order of (18) with D replaced by a moment scale. Proposition A.8 needs no support bound at all: it uses only the log-Sobolev constants, $\alpha_i = \sigma^2$ for $\mathcal{N}(0, \sigma^2 I_d)$. The smoothing refinement (Theorem B.3) and the tuned rate are proved on bounded supports only.

A.5 Global mixing

Corollary A.13 (Global mixing estimates its own population distribution)**.** *For the reuse generator on i.i.d. data, let π be uniform on K-tuples of distinct indices, $K \ge 2$, $d \ge 3$, $n_\pi \ge n$, and let μ_0 have a density bounded by $f_{\max}$ on a set of diameter D. Then $\mathbb{E}W_1(\widehat{\nu}_X, \mathcal{S}_K(\mu_0)) \le C\, D\, n^{-1/(d-\bar{\kappa})}\ell_n$ for $n \ge n_0$, with C, n_0 depending on d, K and $f_{\max}D^d$: $n^{-1/(d-1)}$ for mixup, and $n^{-1/2}$ up to ℓ_n from $K = d - 1$ on.*

This beats $n^{-1/d}$, but towards $\mathcal{S}_K(\mu_0)$, not μ_0: without locality the bias stays.

B THE SMOOTHING REFINEMENT AND THE ROLE OF ARITY

B.1 LOCALITY, SMOOTHING, AND THE RESOLUTION SCALES

Convex mixing does more than reweight: it convolves. Each component $\rho_{\alpha,X} := \mathcal{L}(Y_\alpha(X, u^\alpha) \mid X)$ is the push-forward of the Dirichlet distribution under $u \mapsto \sum_j u_j X_{\alpha_j}$, hence is carried by the $(K-1)$-simplex spanned by $\{X_{\alpha_j}\}_j$, and smooths at the scale of that simplex.

Definition B.1 (Combination-locality scale)**.** For $\alpha \in A$ let $\delta_\alpha(X) := \max_{j,j'} \|X_{\alpha_j} - X_{\alpha_{j'}}\|_2$ be the spread of the combined points, and put $h_\star := \operatorname{ess\,sup}_{\alpha\in A} \delta_\alpha(X)$.

Every K-clique of the ε-graph has spread $< \varepsilon$, so $h_\star \le \varepsilon$; global mixup has $h_\star = \Theta(D)$.

Assumption 2 (κ-dimensional smoothing, second-moment form)**.** Let Y_α, Y'_α be built from the same tuple with independent Dirichlet weights. There is $C_{K,d} \ge 1$ such that for every dyadic partition $\mathcal{Q}_\ell$ into cubes of side s,

$$\sum_{Q\in\mathcal{Q}_\ell} \mathbb{E}\big[\rho_{\alpha,X}(Q)^2\big] \;=\; \mathbb{P}\big(Y_\alpha, Y'_\alpha \text{ fall in one cell}\big) \;\le\; C_{K,d}\Big(\frac{s}{h_\star}\wedge 1\Big)^{\kappa}, \qquad \kappa := \min(K-1, d-1). \tag{24}$$

The average over X is essential: no pointwise version holds, since a tuple with all K points in Q gives $\rho_{\alpha,X}(Q) = 1$ while any right-hand side vanishes with $h_\star$ (Remark C.1). It holds for $K = 2$ with a uniform partner and no hypothesis on μ_0 (Proposition C.3), for global tuples of a bounded density with $h_\star = D$ (Corollary A.13), for k-NN SMOTE and the ε-graph on i.i.d. data with $K \le d$, degenerate tuples included, with a deterministic $h_\star$ (Proposition C.4 and Corollary C.6), and for their split versions at every arity (Lemma E.2). The exponent is the simplex dimension capped at $d-1$: at $K = d+1$ the full d fails by a logarithm (Remark C.7), and the cap changes no rate below, which see κ only through $\min(\kappa, d-2)$ and the sign of $\kappa - (d-2)$. The fluctuation bound is then expressed through the following scales.

Definition B.2 (Truncated entropy integral and resolution scales)**.** For $\theta \in (0, 1]$ set

$$\Phi_d(\theta) \;:=\; \int_\theta^1 t^{-d/2}\,dt \;=\; \begin{cases} 2(1-\sqrt{\theta}), & d = 1,\\ \log\frac{1}{\theta}, & d = 2,\\ \frac{2}{d-2}\big(\theta^{1-d/2} - 1\big), & d \ge 3, \end{cases} \tag{25}$$

and, with $C_Q := \sqrt{d}$, the diameter of the unit cube (any fixed $C_Q \ge 1$ changes only constants, which enter the crossover (84) as C_Q^{2d}), define the *smoothing scale*, the *sampling scale*, the *effective resolution* and the *sub-resolution term*

$$\theta_{\mathrm{sm}} := \frac{h_\star}{C_Q D}, \qquad \theta_{\mathrm{samp}} := S^{1/d}, \qquad \boxed{\theta_\star := \max\{\theta_{\mathrm{sm}}, \theta_{\mathrm{samp}}\}} \tag{26}$$

$$R_{d,\kappa}(\theta_\star) \;:=\; \inf_{0<\theta\le\theta_\star} \Big\{\theta + \sqrt{S}\int_\theta^{\theta_\star} \Big(\frac{t}{\theta_{\mathrm{sm}}}\wedge 1\Big)^{\kappa/2} t^{-d/2}\,dt\Big\}, \qquad R_{d,\kappa}(\theta_{\mathrm{samp}}) \le \theta_{\mathrm{samp}}. \tag{27}$$

In the notation of Section 3, θ_{sm} is t_{sm} and, at $S = 1/n$, θ_{samp} is the sampling scale.

B.2 THE SMOOTHING-REFINED BOUND

The smoothing refinement. Below the locality scale $h_\star$ of Definition B.1 cell masses fluctuate less, so dyadic chaining, which pays $\sqrt{d}D2^{-L}$ for its finest cells, can run further. Assumption 2 enters only through the cell-mass variance (D1) (Section C), which it controls at every scale, below the sampling scale included. Chaining may therefore run past that scale while keeping its finest-cell term, and no coupling hypothesis at the finest level is needed. The scale $\theta_\star$ of Definition B.2 is free of κ, since chaining resolves neither below $h_\star$ nor below $DS^{1/d}$ and the coarser wins, while κ sets how fast the levels below it die: Assumption 2 suppresses a cell-mass variance by $(s/h_\star)^\kappa$, so the levels below $\theta_\star$ contribute (27), where the finest scale θ is free and paid for at its own size.

Theorem B.3 (Smoothing-refined W_1 bound)**.** *Under the hypotheses of Theorem A.7 and Assumption 2,*

$$\mathbb{E}\, W_1(\widehat{\nu}_X, \bar{\nu}) \;\le\; c_d D\big(\sqrt{S}\,\Phi_d(\theta_\star) + R_{d,\kappa}(\theta_\star)\big) \;\wedge\; C_d D\,\psi_d(S), \tag{28}$$

where c_d depends on d, and on K through b_K and $C_{K,d}$; the minimum is termwise. Consequently, with probability at least $1-\delta$,

$$W_1(\widehat{\nu}_X, \mu_0) \;\le\; \beta_1 \;+\; c_d D\big(\sqrt{S}\,\Phi_d(\theta_\star) + R_{d,\kappa}(\theta_\star)\big) \;+\; \sqrt{2V\log(1/\delta)}. \tag{29}$$

If $\theta_\star = \theta_{\mathrm{samp}}$ the refinement reduces to (19), so smoothing lowers the bound only if

$$h_\star \;>\; C_Q D\, S^{1/d} \tag{30}$$

(Proposition B.5), a condition free of κ: blurring below the statistical resolution $DS^{1/d}$ is invisible, however many points are combined.

Proposition B.4 (How much of the gain survives: the role of arity)**.** *Let $d \ge 3$ and $\theta_{\mathrm{sm}} > \theta_{\mathrm{samp}}$. The size of (27) is decided by the sign of $\kappa - (d-2)$. If $\kappa > d-2$, equivalently $K \ge d$, then $R_{d,\kappa} = O(\sqrt{S}\Phi_d(\theta_\star))$ and the full gain over the baseline,*

$$\frac{\Phi_d(\theta_{\mathrm{samp}})}{\Phi_d(\theta_\star)} \;\asymp\; \Big(\frac{\theta_\star}{\theta_{\mathrm{samp}}}\Big)^{d/2-1} = \Big(\frac{h_\star}{C_Q D S^{1/d}}\Big)^{d/2-1}, \tag{31}$$

is realised. If $\kappa = d-2$, i.e. $K = d-1$, it survives up to a factor $\log(\theta_\star/\theta_{\mathrm{samp}})$. If $\kappa < d-2$, which covers mixup ($K = 2$) in every dimension $d \ge 4$, $R_{d,\kappa}$ dominates and the bound is the baseline times $(\theta_{\mathrm{samp}}/\theta_\star)^{\kappa/(d-\kappa)} = (C_Q D S^{1/d}/h_\star)^{\kappa/(d-\kappa)}$, a factor that is sharp when about n distinct tuples are drawn (Proposition C.10). Section B carries the explicit constants. For $d \le 2$ every κ is super-critical and for $d = 3$ pairs are exactly critical, so in low dimension the bound for mixup already has the full gain, up to that logarithm at $d = 3$.

B.3 Proof of the smoothing refinement

The baseline member $C_d D\psi_d(S)$ of the minimum in (28) is (18), proved in Section A.4. We prove the refined member, then the following proposition and Proposition B.4.

Proposition B.5 (Never worse, and when smoothing can help at all)**.** *If $\theta_\star = \theta_{\mathrm{samp}}$ then $R_{d,\kappa} \le \theta_{\mathrm{samp}}$ and (29) reduces to (19) in every dimension; since Φ_d is strictly decreasing, smoothing lowers the bound only if $\theta_{\mathrm{sm}} > \theta_{\mathrm{samp}}$, i.e. iff (30) holds, a condition free of κ: blurring at a scale finer than the statistical resolution $DS^{1/d}$ is invisible to the bound, however many points are combined.*

Step 1: the level decomposition. Run the dyadic estimate of Section A.4 Step 3 for the pair $(\widehat{\nu}_X, \bar{\nu})$, writing $\theta_\ell := 2^{-\ell}$ for the relative scale and $s_\ell = D\theta_\ell$ for the side of a cell at level ℓ. By (D1) and Jensen's inequality,

$$\mathbb{E}|\Delta_Q| \;\le\; \sqrt{\mathrm{Var}(\widehat{\nu}_X(Q))} \;\le\; \sqrt{b_K S}\,\Big(\frac{\theta_\ell}{\theta_{\mathrm{sm}}} \wedge 1\Big)^{\kappa/2}\sqrt{\bar{\nu}(Q)}. \tag{32}$$

Only independence enters here, so the scale is $\sqrt{S}$ and not $\sqrt{V_1}$: $\mathbf{1}_Q$ is not Lipschitz, and log-Sobolev constants do not control the variance of a cell mass. If μ_i is uniform on a segment of length ε crossing ∂Q, then $\alpha_i \asymp \varepsilon^2$ while $\mathrm{Var}\,\mathbf{1}_Q(X_i) = \frac{1}{4}$, so $\mathrm{Var}\,\mu_w(Q) = S/4 \gg V_1/4$. The deviation proxy, V or V_1, enters only the deviation term of (29), through the Lipschitz map of Section A.4 Step 8. Summing over $Q \in \mathcal{Q}_\ell$ and applying Cauchy–Schwarz with $\sum_Q \bar{\nu}(Q) = 1$ and $\#\mathcal{Q}_\ell = \theta_\ell^{-d}$,

$$\mathbb{E}A_\ell \;\le\; \sqrt{b_K S}\,\Big(\frac{\theta_\ell}{\theta_{\mathrm{sm}}} \wedge 1\Big)^{\kappa/2}\theta_\ell^{-d/2}. \tag{33}$$

The transport weight at level ℓ is $2^{-(\ell-1)} \le 2\theta_\ell$. Since Proposition C.2 gives (33) for *every* dyadic partition, the finer ones included, the recursion (22) may be run down to any level L, paying $\sqrt{d}D\theta_L$ for the within-cell matching there; nothing is assumed about that finest coupling. For every $L \ge 1$, therefore,

$$\mathbb{E}W_1(\widehat{\nu}_X, \bar{\nu}) \;\le\; \sqrt{d}D\theta_L + c_d D\sqrt{S}\sum_{\ell=1}^{L}\Big(\frac{\theta_\ell}{\theta_{\mathrm{sm}}} \wedge 1\Big)^{\kappa/2}\theta_\ell^{1-d/2}. \tag{34}$$

Stopping at $\theta_L = \theta_{\mathrm{samp}}$ recovers the baseline, since the finest term is then $\sqrt{d}DS^{1/d}$; the refinement comes from stopping lower.

Step 2: splitting the sum at $\theta_\star$. Split (34) at $\theta_\star = \max\{\theta_{\mathrm{sm}}, \theta_{\mathrm{samp}}\}$.

Levels $\theta_\ell \geq \theta_\star$. Here $\theta_\ell \geq \theta_{\mathrm{sm}}$, the minimum equals 1, and the substitution $t = 2^{-\ell}$, for which $t^{1-d/2} = 2^{\ell(d/2-1)}$ and dt/t matches the dyadic spacing, converts the sum into the truncated entropy integral:

$$\sum_{\theta_\ell \geq \theta_\star} \theta_\ell^{1-d/2} \asymp \int_{\theta_\star}^{1} t^{-d/2}\, dt = \Phi_d(\theta_\star). \tag{35}$$

Levels $\theta_L \leq \theta_\ell < \theta_\star$, and the finest term. The same substitution gives

$$\sqrt{d}\theta_L + c_d\sqrt{S} \sum_{\theta_L \leq \theta_\ell < \theta_\star} \Big(\frac{\theta_\ell}{\theta_{\mathrm{sm}}} \wedge 1\Big)^{\kappa/2} \theta_\ell^{1-d/2} \lesssim \theta_L + \sqrt{S} \int_{\theta_L}^{\theta_\star} \Big(\frac{t}{\theta_{\mathrm{sm}}} \wedge 1\Big)^{\kappa/2} t^{-d/2}\, dt. \tag{36}$$

Every $\theta \in (0, \theta_\star]$ lies within a factor 2 of a dyadic θ_L, so minimising over L gives $R_{d,\kappa}(\theta_\star)$ of (27) up to a constant absorbed into c_d. Adding the two pieces proves the refined member of (28). No hypothesis beyond Theorem A.7 and Assumption 2 has been used, in any regime of κ. In particular $h_\star$ enters only through Assumption 2, so it may be any deterministic scale for which that assumption holds, such as the k-NN scale of Proposition C.4. The spread bound of Definition B.1 matters only for β_1 in (29), and for k-NN neighbourhoods it holds with probability $1 - \delta$ (Lemma G.2), so there (29) holds with probability $1 - 2\delta$.

Step 3: never worse, and the threshold (Proposition B.5). If $\theta_\star = \theta_{\mathrm{samp}} = S^{1/d}$ then $R_{d,\kappa} \leq \theta_{\mathrm{samp}} \leq \psi_d(S)$ (take $\theta = \theta_\star$) and one checks directly from (25), for $d = 1$, $\theta_\star = S$ and $\Phi_1(S) = 2(1 - \sqrt{S}) \asymp 2$, so $\sqrt{S}\Phi_1 \asymp \sqrt{S} = \psi_1(S)$; for $d = 2$, $\theta_\star = S^{1/2}$ and $\Phi_2 = \frac{1}{2}\log\frac{1}{S}$, so $\sqrt{S}\Phi_2 \asymp \sqrt{S}\log\frac{1}{S} = \psi_2(S)$; for $d \geq 3$, $\theta_\star = S^{1/d}$ and $\theta_\star^{1-d/2} = S^{\frac{1}{d}-\frac{1}{2}}$, so $\sqrt{S}\Phi_d \asymp \sqrt{S} \cdot S^{\frac{1}{d}-\frac{1}{2}} = S^{1/d} = \psi_d(S)$. Since Φ_d is strictly decreasing and $R_{d,\kappa}$ is of baseline order in that case, the refined member can be below the baseline only if $\theta_{\mathrm{sm}} > \theta_{\mathrm{samp}}$, i.e. $h_\star/(C_Q D) > S^{1/d}$, which is (30). No exponent in κ appears in the comparison, since both scales are relative lengths.

Step 4: the three arity regimes (Proposition B.4). Assume $d \geq 3$ and $\theta_{\mathrm{sm}} > \theta_{\mathrm{samp}}$, so $\theta_\star = \theta_{\mathrm{sm}}$ and $\sqrt{S} = \theta_{\mathrm{samp}}^{d/2}$; put $p := (\kappa - d)/2$ and $\Lambda := \sqrt{S}\,\theta_{\mathrm{sm}}^{-\kappa/2}$. For $\theta \leq \theta_{\mathrm{sm}}$ the integral in (27) is $\theta_{\mathrm{sm}}^{-\kappa/2}\int_\theta^{\theta_{\mathrm{sm}}} t^p\, dt$. If $p > -1$, i.e. $\kappa > d - 2$, let $\theta \to 0$: the integral converges and $R_{d,\kappa} \leq \frac{2}{\kappa - d + 2}\sqrt{S}\,\theta_{\mathrm{sm}}^{1-d/2} \leq \frac{d-2}{\kappa-d+2}\sqrt{S}\,\Phi_d(\theta_\star)(1 + o(1))$, which is case (i), and the gain (31) follows from $\Phi_d(\theta) \asymp \frac{2}{d-2}\theta^{1-d/2}$. If $p = -1$, i.e. $\kappa = d - 2$, the integral is $\theta_{\mathrm{sm}}^{1-d/2}\log(\theta_{\mathrm{sm}}/\theta)$; taking $\theta = T := \sqrt{S}\,\theta_{\mathrm{sm}}^{1-d/2}$, which is at most θ_{sm} up to a constant since $T/\theta_{\mathrm{sm}} \asymp (\theta_{\mathrm{samp}}/\theta_{\mathrm{sm}})^{d/2}$, gives $R_{d,\kappa} \leq T\big(1 + \frac{d}{2}\log(\theta_{\mathrm{sm}}/\theta_{\mathrm{samp}}) + O(1)\big)$ with $T \asymp \sqrt{S}\Phi_d(\theta_\star)$, which is case (ii). If $p < -1$, i.e. $\kappa < d - 2$, the integral is at most $\frac{2}{d-2-\kappa}\theta_{\mathrm{sm}}^{-\kappa/2}\theta^{-(d-2-\kappa)/2}$, and $\theta + \frac{2\Lambda}{d-2-\kappa}\theta^{-(d-2-\kappa)/2}$ is minimised at $\theta = \Lambda^{2/(d-\kappa)}$, with value $\frac{d-\kappa}{d-2-\kappa}\Lambda^{2/(d-\kappa)}$. Since $\Lambda^{2/(d-\kappa)} \asymp \theta_{\mathrm{samp}}^{d/(d-\kappa)}\theta_{\mathrm{sm}}^{-\kappa/(d-\kappa)}$ and the minimiser lies below θ_{sm},

$$R_{d,\kappa}(\theta_\star) \leq \frac{d-\kappa}{d-2-\kappa}\, c\,\theta_{\mathrm{samp}}\Big(\frac{\theta_{\mathrm{samp}}}{\theta_\star}\Big)^{\kappa/(d-\kappa)}, \tag{37}$$

while $\sqrt{S}\Phi_d(\theta_\star) \asymp \theta_{\mathrm{samp}}(\theta_{\mathrm{samp}}/\theta_\star)^{d/2-1}$ is smaller, because $(d-2)(d-\kappa) > 2(d-2) > 2\kappa$. As the baseline is $\sqrt{S}\Phi_d(\theta_{\mathrm{samp}}) \asymp \theta_{\mathrm{samp}}$, this is case (iii). With $n_\Sigma \asymp n$ distinct tuples the covering radius $r_\star$ of Proposition C.10 equals the right-hand side of (37) up to constants, so (iii) is sharp. Finally $\kappa = \min(K - 1, d - 1) > d - 2$ holds iff $K \geq d$, and $\kappa = d - 2$ iff $K = d - 1$.

Step 5: assembling. Combine (1), the expectation bound (28) and the concentration of Section A.4 Step 8:

$$W_1(\widehat{\nu}_X, \mu_0) \leq \beta_1 + H(X) \leq \beta_1 + \mathbb{E}H + \sqrt{2V\log(1/\delta)}$$

with probability at least $1 - \delta$, which is (29). Under log-Sobolev inequalities with $\alpha_i \leq \alpha$ the deviation term may be replaced by $\sqrt{2\alpha S\log(1/\delta)}$. If moreover $h_\star \leq C_Q D S^{1/d}$, $\theta_\star$ collapses to $S^{1/d}$, $R_{d,\kappa} \leq \theta_\star$, and by Step 3 $c_d D(\sqrt{S}\Phi_d(\theta_\star) + R_{d,\kappa}) \lesssim D\psi_d(S)$, so (29) reduces to (19). □

C ASSUMPTION 2: WHAT IT GIVES, WHERE IT HOLDS, AND SHARPNESS

The proof consumes Assumption 2 only through one consequence, which we record because it, and not a pointwise bound, is what is used. Writing $s = D\theta$ and $\theta_{\mathrm{sm}} := h_\star/(C_Q D)$, there is $b_K \geq 1$ with

$$\text{(D1)} \quad \mathrm{Var}\left(\widehat{\nu}_X(Q)\right) \;\leq\; b_K\, S\left(\frac{\theta}{\theta_{\mathrm{sm}}} \wedge 1\right)^{\kappa} \bar{\nu}(Q) \quad \text{for every cube } Q \text{ of side } s, \tag{38}$$

(D1) is the bounded-range variance estimate with the overlap structure of Lemma A.6, and for $\theta \geq \theta_{\mathrm{sm}}$ is the unsmoothed bound of Section A.4 Step 7. It holds at every scale θ, below θ_{samp} included, which is what lets Section B run the chaining to an optimised finest level and keep the finest-cell term $\sqrt{d}D\theta_L$ rather than assume it away. The alternative would be a finest-level coupling, (D2): that at the finest retained level the restrictions of $\widehat{\nu}_X$ and $\bar{\nu}$ to a cell couple at the same variance-controlled cost, so that no separate term $\sqrt{d}D\theta$ arises. No result below uses (D2). Its conclusion is reached without it for $K \geq d$ (Proposition C.8, a chaining-free route to case (i) of Proposition B.4), and below the threshold (D2) fails (Proposition C.10), which is what makes case (iii) sharp.

Write $T_\alpha := \mathrm{conv}\{X_{\alpha_j}\}_j$ for the simplex of a tuple, and call it $h_\star$-regular if it contains a ball of radius $h_\star/c_{\mathrm{reg}}$ and is contained in one of radius $h_\star$.

Remark C.1 (Why the condition must be an average). A pointwise version of Assumption 2, meaning a bound on $\rho_{\alpha,X}(Q)$ valid for every realisation, cannot hold for a local generator, in either of the two shapes one might write. Suppose first that the right-hand side carries the mean measure, as in $\rho_{\alpha,X}(Q) \leq C_{K,d}(s/h_\star \wedge 1)^\kappa \bar{\nu}_\alpha(Q^{\oplus h_\star})$. Take a realisation in which all K points of the tuple lie in a cube Q of side $s \asymp h_\star$: the simplex T_α is then contained in Q by convexity, so the left-hand side is 1, while $\bar{\nu}_\alpha(Q^{\oplus h_\star}) \leq f_{\max}(3s)^d \to 0$ as $h_\star \to 0$. No constant free of $h_\star$ and μ_0 survives, and the configuration has positive probability; it is in fact the typical one for an $h_\star$-local generator. Dropping the mean-measure factor repairs the scaling but not the statement: $\rho_{\alpha,X}(Q) \leq C_{K,d}(s/h_\star)^\kappa$ requires every drawn simplex to be $h_\star$-regular, and random tuples are degenerate with positive probability: two k-NN partners at distance $\varepsilon h_\star$ occur with probability $\asymp \varepsilon^d$, and the component then concentrates at a scale far below $h_\star$.

What the proof consumes is neither: only the summed second moment of Assumption 2, through $\mathrm{Cov}(Z_\alpha, Z_\beta) \leq \frac{1}{2}(\mathbb{E}Z_\alpha^2 + \mathbb{E}Z_\beta^2)$ for $Z_\alpha := \rho_{\alpha,X}(Q)$. Averaging lets a degenerate tuple contribute in proportion to its probability, which is what makes the condition attainable.

Proposition C.2 (Assumption 2 implies (D1) in summed form)**.** *Two components are independent unless their tuples share a data index, and* $\#\{\beta : \beta \sim \alpha\} \leq \sum_{j\leq K} r_{\alpha_j}$ *with* $\sum_\alpha \sum_{j\leq K} r_{\alpha_j} = \sum_i r_i^2 = (n_\pi K)^2 S$. *Hence*

$$\sum_{Q\in\mathcal{Q}_\ell} \mathrm{Var}\left(\widehat{\nu}_X(Q)\right) \;\leq\; \frac{1}{{n_\pi}^2} \sum_Q \sum_\alpha \mathbb{E}[Z_\alpha^2]\, \#\{\beta : \beta \sim \alpha\} \;\leq\; K^2 S\, C_{K,d}\Big(\frac{s}{h_\star} \wedge 1\Big)^\kappa,$$

and Cauchy–Schwarz over the θ^{-d} *cells gives* $\mathbb{E}A_\ell \leq \theta^{-d/2}\big(K^2 S C_{K,d}(\theta/\theta_{\mathrm{sm}} \wedge 1)^\kappa\big)^{1/2}$, *which is* (33)*, the only form in which (D1) is used in Section B.*

Proposition C.3 (Assumption 2 for pairs, unconditionally)**.** *Let* $K = 2$*:* $Z_1 \sim \mu_0$, Z_2 *uniform on* $B(Z_1, h_\star)$, *and* $Y = UZ_1 + (1-U)Z_2$ *with* $U \sim \mathrm{Unif}[0,1]$. *Then for every grid of side* $s \leq h_\star$,

$$\mathbb{P}\big(Y, Y' \text{ in one cell}\big) \;\leq\; \frac{2d\sqrt{d}}{d-1} \cdot \frac{s}{h_\star}, \tag{39}$$

so Assumption 2 holds with $\kappa = 1$ *and* $C_{2,d} \leq 2d\sqrt{d}/(d-1)$, *for every* $d \geq 2$ *and with no hypothesis on* μ_0. *If instead* $Z_2 \sim \mu_0|_{B(Z_1,h_\star)}$ *under Assumption 1, the bound holds with the extra factor* c_f.

Proof. Conditionally on (Z_1, Z_2), Y and Y' are independent and uniform on the segment of length $L := \|Z_1 - Z_2\|$, so they share a cell only if $\|Y - Y'\|_\infty \leq s$, i.e. $|U - U'|L \leq \sqrt{d}s$; as $U - U'$ has density at most 1, the conditional probability is at most $\min(1, 2\sqrt{d}s/L)$. The radius of a uniform draw in $B(Z_1, h_\star)$ has density $dL^{d-1}/h_\star^d$ on $(0, h_\star)$, so $\mathbb{E}[2\sqrt{d}s/L] = 2\sqrt{d}s\,\frac{d}{h_\star^d}\int_0^{h_\star} L^{d-2}dL = \frac{2d\sqrt{d}}{d-1}\frac{s}{h_\star}$, and the cap only lowers it. For $Z_2 \sim \mu_0|_{B(Z_1,h_\star)}$ the density of Z_2 is at most $\|f\|_\infty/(c_0 h_\star^d) = c_f/(\omega_d h_\star^d)$, which is c_f times the uniform one. □

Proposition C.4 (Assumption 2 for k-NN SMOTE, degenerate tuples included)**.** *Let $X_1, \dots, X_n$ be i.i.d. μ_0 with density $f \le f_{\max}$ and $\mu_0(B(x,\rho)) \ge c_0\rho^d$ for every $x \in \operatorname{supp}\mu_0$ and $\rho \le D$, as in Lemma G.2. Let π draw a base uniformly from the sample and $K-1$ distinct partners uniformly among its k nearest neighbours, with $2 \le K \le d$ and $k \ge \log(n/\delta)$, and let $h_\star := 2\big(\frac{2k+8\log(n/\delta)}{c_0(n-1)}\big)^{1/d}$, twice the right-hand side of (80). This $h_\star$ is deterministic. With probability at least $1-\delta$ it bounds the spread of every tuple, so it is an admissible locality scale in Definition B.1 on that event; and Assumption 2 holds with it* unconditionally*, with $\kappa = K-1$ and*

$$C_{K,d} \;\le\; (K-1)!\,\omega_{K-1}d^{(K-1)/2}\,c\,\Lambda_{K,d}, \qquad \Lambda_{K,d} := \prod_{j=2}^{K}\Big(1+\frac{C_0}{d-j+1}\Big), \tag{40}$$

where $C_0 := c_d f_{\max}/c_0$ and ω_{K-1} is the volume of the unit ball in $\mathbb{R}^{K-1}$. No regularity is assumed of the drawn simplex: degenerate tuples are included, and contribute in proportion to their probability.

Proof. Step 1: reduction to a small ball. With $T_j := Z_j - Z_1$ and $M := [T_2, \dots, T_K] \in \mathbb{R}^{d\times(K-1)}$, the two synthetic points of a tuple differ by $Y - Y' = \sum_{j\ge2}(u_j - u'_j)T_j = Mv$, since $\sum_j(u_j - u'_j) = 0$. The vector $(u_2, \dots, u_K)$ is uniform on $\{u \ge 0, \sum u_j \le 1\}$ with density $(K-1)!$, so v, a difference of two independent copies, has density at most $(K-1)!$ by the convolution bound. Sharing a cell of side s forces $\|Mv\|_\infty < s$, hence $\|Mv\|_2 < \sqrt{d}s$, and the preimage of that ball is an ellipsoid of volume $\omega_{K-1}(\sqrt{d}s)^{K-1}/\prod_i \sigma_i(M)$. Therefore

$$\mathbb{P}\big(Y, Y' \text{ in one cell} \mid M\big) \;\le\; 1 \wedge \frac{(K-1)!\,\omega_{K-1}d^{(K-1)/2}\,s^{K-1}}{\prod_i \sigma_i(M)}. \tag{41}$$

Step 2: the neighbours are conditionally i.i.d. Condition on the base $X_b = x$ and on ρ, the distance from x to its $(k+1)$-st nearest neighbour. Given a region and the number of sample points in it, those points are i.i.d. with μ_0 restricted to the region; hence the k nearest neighbours of x are i.i.d. $\mu_0|_{B(x,\rho)}$. A uniformly chosen $(K-1)$-subset of i.i.d. variables is again i.i.d., so $T_2, \dots, T_K$ are i.i.d. draws from $\mu_0(\cdot + x)|_{B(0,\rho)}$. Conditioning on the $(k+1)$-st radius rather than the k-th is what keeps every partner strictly inside the ball, so no boundary point has to be treated separately.

Step 3: peeling, with no regularity. Gram–Schmidt gives $\prod_i \sigma_i(M) = \prod_{j=2}^K d_j$ with $d_j := \operatorname{dist}\big(T_j, \operatorname{span}(T_2, \dots, T_{j-1})\big)$. Given the past, T_j is an independent draw and the span has dimension at most $j-2$, so the set of points within $\delta\rho$ of it meets $B(0,\rho)$ in volume at most $c_d\rho^d\delta^{\,d-j+2}$, while $\mu_0(B(x,\rho)) \ge c_0\rho^d$. Hence

$$\mathbb{P}\big(d_j \le \delta\rho \mid T_2, \dots, T_{j-1}\big) \;\le\; C_0\,\delta^{\,\gamma_j}, \qquad \gamma_j := d-j+2 \;\ge\; d-K+2 \;\ge\; 2,$$

the last inequality because $K \le d$. A non-negative variable with $\mathbb{P}(W \le \delta) \le C_0\delta^\gamma$ and $\gamma > 1$ has $\mathbb{E}[1/W] = \int_0^\infty \mathbb{P}(W < 1/t)\,dt \le 1 + C_0/(\gamma-1)$, so $\mathbb{E}[\rho/d_j \mid \text{past}] \le 1 + C_0/(d-j+1)$ *uniformly in the past*. The tower property applied from $j = K$ down to $j = 2$ gives

$$\mathbb{E}\Big[\frac{\rho^{K-1}}{\prod_i \sigma_i(M)} \;\Big|\; x, \rho\Big] \;=\; \mathbb{E}\Big[\prod_{j=2}^{K}\frac{\rho}{d_j} \;\Big|\; x, \rho\Big] \;\le\; \Lambda_{K,d}. \tag{42}$$

This is the step that tolerates degeneracy: a near-flat simplex makes one factor large, but only on an event whose probability decays as δ^{γ_j}, and the two balance because $\gamma_j > 1$.

Step 4: the neighbourhood radius. By the Chernoff bound for $\mathrm{Bin}(n-1, \mu_0(B(x, \lambda(k/n)^{1/d})))$, the number of points in a ball of radius $\lambda(k/n)^{1/d}$ exceeds k with probability at most $(ec_1\lambda^d)^k$, so $\mathbb{P}(\rho \le \lambda(k/n)^{1/d}) \le (ec_1\lambda^d)^k$ and the negative moment $\mathbb{E}[\rho^{-(K-1)}]$ converges whenever $dk > K-1$, with $\mathbb{E}[(h_\star/\rho)^{K-1}] \le c$ once $k \ge \log(n/\delta)$, since then $h_\star \le c'(k/n)^{1/d}$. Only this lower tail of ρ enters, with no event excluded, which is why Assumption 2 holds for the deterministic $h_\star$ without a probability qualifier; Lemma G.2 is needed only for the upper bound on the spread, that is for β_1. Combining with (41) and (42),

$$\mathbb{P}\big(Y, Y' \text{ in one cell}\big) \;\le\; (K-1)!\,\omega_{K-1}d^{\frac{K-1}{2}}\Big(\frac{s}{h_\star}\Big)^{K-1}\mathbb{E}\Big[\Big(\frac{h_\star}{\rho}\Big)^{K-1}\frac{\rho^{K-1}}{\prod_i\sigma_i}\Big] \;\le\; C_{K,d}\Big(\frac{s}{h_\star}\Big)^{K-1},$$

which is (40); for $s \ge h_\star$ the claim is trivial since the probability is at most one. □

Proposition C.5 (One estimate, both hypotheses)**.** *The proof of Proposition C.4 delivers, before any reference to cells, the* small-ball bound

$$\mathbb{P}\big(\|Y_\alpha - Y'_\alpha\|_2 \le t\big) \;\le\; C_{K,d}\Big(\frac{t}{h_\star} \wedge 1\Big)^{\kappa} \qquad \textit{for all } t > 0, \tag{43}$$

with the constant of (40). Two things follow from (43) alone. First, Assumption 2, since a shared cell of side s forces $\|Y_\alpha - Y'_\alpha\|_2 < \sqrt{d}s$. Second, for $d \ge 3$ and $\kappa > d-2$, the Fourier hypothesis (46) in the only form Proposition C.8 consumes:

$$\sum_{k\neq 0} \frac{\operatorname{Var}\big(\widehat{\nu_X}(k)\big)}{4\pi^2|k|^2} \;\le\; c_d K^2 S\, C_{K,d}\Big(\frac{1}{\kappa - d + 2} + \frac{1}{d-2}\Big) h_\star^{2-d}, \tag{44}$$

Proposition C.4 covers $2 \le K \le d$, so this derivation reaches the bottom of the range Proposition C.8 needs, $K = d$; above it the exponent $\kappa = d-1$ of Lemma E.2 gives it in the same way for the split generators, as in the proof of Theorem F.1. What it is not is a second, independent hypothesis.

Proof. (43) is Steps 1–4 of Proposition C.4 with $\sqrt{d}s$ renamed t: nothing there used that the target set was a cube. For the second claim, the dependency count of Proposition C.2 applies verbatim to the complex variables $Z_\alpha := \hat\rho_{\alpha,X}(k)$, giving $\operatorname{Var}(\widehat{\nu_X}(k)) \le K^2 S \max_\alpha \mathbb{E}|\hat\rho_{\alpha,X}(k)|^2$. Now the second moment of a component's Fourier coefficient is a two-point function of the same pair that (43) controls:

$$\mathbb{E}\big|\hat\rho_{\alpha,X}(k)\big|^2 = \mathbb{E}_X\Big|\mathbb{E}_u e^{-2\pi i k\cdot Y_\alpha}\Big|^2 = \mathbb{E}\big[e^{-2\pi i k\cdot(Y_\alpha - Y'_\alpha)}\big],$$

because pairing $\hat\rho$ with its conjugate amounts to drawing the Dirichlet weights twice. Summing against $4\pi^2|k|^2$ and exchanging the sum with the expectation, $\sum_{k\neq0}\mathbb{E}|\hat\rho_\alpha(k)|^2/(4\pi^2|k|^2) = \mathbb{E}[G(Y_\alpha - Y'_\alpha)]$, where G is the mean-zero Green's function of $\mathbb{T}^d$, which satisfies $G(x) \le c_d\|x\|^{2-d} + c'_d$ for $d \ge 3$. Finally, for a non-negative W, $\mathbb{E}[W^{2-d}] = (d-2)\int_0^\infty t^{1-d}\mathbb{P}(W \le t)\,dt$, so by (43)

$$\mathbb{E}\big[\|Y_\alpha - Y'_\alpha\|^{2-d}\big] \;\le\; (d-2)\Big[C_{K,d}h_\star^{-\kappa}\int_0^{h_\star} t^{\,1-d+\kappa}dt + \int_{h_\star}^{\infty} t^{\,1-d}dt\Big] \;=\; C_{K,d}\,h_\star^{2-d}\Big[\tfrac{d-2}{\kappa-d+2} + 1\Big],$$

the first integral converging at the origin when $\kappa > d-2$, and only then. □

Corollary C.6 (The ε-graph generator)**.** *Let π draw a base uniformly and $K-1$ distinct partners uniformly among the sample points within distance ε of it, conditioned on the base having degree at least $K-1$. Then (43), Assumption 2 and (44) hold with $h_\star = \varepsilon$, the same $\Lambda_{K,d}$, and for every $2 \le K \le d$, with no condition on n or on the degree distribution. Here ε is the partner radius; the spread is at most 2ε, and since a bound at the smaller scale is the stronger one, they hold a fortiori with the $h_\star = 2\varepsilon$ of Definition B.1.*

Proof. Step 2 of Proposition C.4 needs only that, conditionally on the base x and on the number of sample points in $B(x,\varepsilon)$, those points are i.i.d. $\mu_0|_{B(x,\varepsilon)}$, which is the same fact about a region and its count, and a uniformly chosen $(K-1)$-subset of them is again i.i.d.; no $(k+1)$-st radius is needed, because the ball is prescribed rather than read off the sample. Step 3 is unchanged, with ρ replaced by ε. Step 4 disappears: the radius is deterministic, so the negative moment $\mathbb{E}[(h_\star/\rho)^{K-1}]$ that it controlled equals one. The degree conditioning enters only through the choice of base and leaves the conditional distribution of the partners untouched. □

Between Proposition C.3, Proposition C.4 and Corollary C.6, Assumption 2 is a theorem for pairs at any locality, and for both neighbourhood graphs at every arity $2 \le K \le d$, degenerate tuples included. Above that its exponent is $\kappa = d-1$, which holds for every $K > d$ by aggregation onto a d-point sub-tuple (Lemma E.2), while the full exponent d acquires a logarithm at $K = d+1$ (Remark C.7); Proposition C.8 asks only for (44), and $\kappa = d-1 > d-2$ holds there outright.

Remark C.7 (What happens at $K = d+1$, and what it costs). Step 3 is the only place $K \le d$ is used, through $\gamma_K = d - K + 2 > 1$. At $K = d+1$ the last factor has $\gamma_K = 1$ and $\mathbb{E}[\rho/d_K]$ diverges logarithmically; truncating at the scale $\delta = s/\rho$, below which (41) saturates at one, replaces $\Lambda_{K,d}$ by $O(\log(h_\star/s))$, and the full exponent d holds only with the extra logarithm. Simulation of the generator itself agrees: with $n = 800$ and $k = 24$, the mean of $\rho^{K-1}/\prod_i \sigma_i$ is stable in the sample

size for $K \le d$ ($1.38, 2.49, 7.23$ at $d = 4$ for $K = 2, 3, 4$ and $1.55, 3.95$ at $d = 8$ for $K = 3, 5$), and creeps upward with it at $K = d + 1$ ($91 \to 154 \to 196$ over a sixteen-fold increase), as a divergent mean does. So the boundary of Proposition C.4 is the generator's rather than the proof's, and it is the same $K \le d$ boundary that Proposition B.4, Remark C.9 and (45) reach from three other directions. It is a boundary for the full exponent only: aggregating the Dirichlet weights onto d of the K points costs one dimension and removes the divergence, since the aggregate weight $B \sim \mathrm{Beta}(d, K - d)$ has $\mathbb{E}B^{-(d-1)} = \binom{K-1}{d-1} < \infty$, and the exponent $d - 1 > d - 2$ that remains, the κ of Assumption 2 above $K = d$, keeps every rate (Lemma E.2).

For $K \ge 3$ the same computation is a small-ball problem. Writing $Y - Y' = \sum_{j \ge 2}(u_j - u'_j)(Z_j - Z_1) = Mv$ with M the $d \times (K - 1)$ matrix of offsets and v of bounded density on $\mathbb{R}^{K-1}$, the conditional probability is $\asymp \min\{1, (s/h_\star)^{K-1}/J\}$ with $J := \prod_i \sigma_i(M)/h_\star^{K-1}$ the normalised simplex volume. Since $\mathbb{P}(J \le \varepsilon) \asymp \varepsilon^{\,d-K+2}$,

$$\mathbb{P}\big(Y, Y' \text{ in one cell}\big) \;\asymp\; c_K\, \mathbb{E}\big[1/J\big] \left(\frac{s}{h_\star}\right)^{K-1}, \tag{45}$$

and $\mathbb{E}[1/J] < \infty$ exactly when $K \le d$, with a logarithmic divergence at $K = d + 1$, which Proposition C.4 makes rigorous for the k-NN generator. The exponent $K - 1$ of (45) survives up to $K = d + 1$, and the constant does not: measured on $h_\star$-local tuples, $\mathbb{E}[1/J]$ is $1.33, 2.26, 6.04$ for $d = 4$ at $K = 2, 3, 4$ and $1.42, 3.02, 47.2$ for $d = 8$ at $K = 3, 5, 8$, all stable in the sample size, while at $K = d + 1$ the estimate creeps upward with it, as a divergent mean does. The measured same-cell ratios follow: they settle at 1.13 and 3.7 for $d = 4$, $K = 2, 3$ as $s/h_\star \downarrow 0$, and do not settle at $K = d + 1$. This is the quantitative content of the $K \ge d$ boundary seen from a third side, after Proposition B.4 and Remark C.9, and a reason to prefer the route of Proposition C.8, whose Fourier hypothesis (46) is a variance and so inherits none of this.

The $\dot{H}^{-1}$ route, and a covering bound below the threshold

For $K \le d$ the simplex T_α has dimension $\kappa = K - 1 < d$, so $\rho_{\alpha,X}$ is singular with respect to Lebesgue measure and there is no synthetic density to compare with that of $\bar{\nu}$ at the finest level. Two standard devices do without one, the quantitative Monge–Ampère linearisation of Goldman et al. (2021) and the heat-semigroup regularisation of Ambrosio et al. (2019), and both pass through the same object, the $\dot{H}^{-1}$ energy of the fluctuation. Carrying that object through gives case (i) of Proposition B.4 without chaining on one side of $\kappa = d - 2$, and on the other shows (D2) to be *false* without a further hypothesis.

Write the generator on the torus $\mathbb{T}^d = [0, 1)^d$ with $\widehat{\mu}(k) = \int e^{-2\pi i k \cdot y}\, d\mu(y)$, and record Assumption 2 on the Fourier side:

$$\text{(D1}') \qquad \mathrm{Var}\left(\widehat{\widehat{\nu}_X(k)}\right) \;\le\; b_K S\Big(1 \wedge \Big(\frac{C_Q}{h_\star |k|}\Big)^{\kappa}\Big), \qquad k \in \mathbb{Z}^d \setminus \{0\}. \tag{46}$$

(D1′) is (D1) read in frequency, and it is not an extra assumption: Proposition C.5 derives the only form of it the proof below consumes, the sum (44), from the small-ball bound (43), hence from Proposition C.4, for both neighbourhood graphs and every $2 \le K \le d$. We state it as a display because that is the form in which it enters the Parseval computation.

Proposition C.8 (A chaining-free route when $K \ge d$)**.** *Let $d \ge 3$, $\kappa > d - 2$, and let $\bar{\nu}$ have a density bounded below by $f_{\min} > 0$ on $\mathbb{T}^d$ (take $D = 1$; both sides below are homogeneous of degree one in D, so the general case follows by rescaling). Under (46),*

$$\mathbb{E}\, W_1(\widehat{\nu}_X, \bar{\nu}) \;\le\; \mathbb{E}\, W_2(\widehat{\nu}_X, \bar{\nu}) \;\le\; 2 f_{\min}^{-1/2} \big(\mathbb{E}\|\widehat{\nu}_X - \bar{\nu}\|^2_{\dot{H}^{-1}}\big)^{1/2} \;\le\; c_{d,K} \sqrt{S}\, \theta_\star^{\,1-d/2}, \tag{47}$$

which is the $\Phi_d(\theta_\star)$ term of Theorem B.3, obtained with no density, no chaining and no finest-cell step. Since $\kappa = \min(K - 1, d - 1) > d - 2$ iff $K \ge d$, this recovers case (i) of Proposition B.4 on the torus, independently of Section B. It needs (46) and $f_{\min} > 0$ and does not cover domains with boundary; Section B does.

Proof. Peyre's inequality (Peyre, 2018) gives $W_2(\mu, \nu) \le 2\|\mu - \nu\|_{\dot{H}^{-1}(\nu)}$ for an arbitrary probability measure μ and a probability measure ν; no density is required of μ, which is what makes it usable

here. Bounding the weighted norm by $f_{\min}^{-1/2}$ times the flat one and using Parseval,

$$\mathbb{E}\|\widehat{\nu}_X - \bar{\nu}\|_{\dot{H}^{-1}}^2 = \sum_{k\neq 0} \frac{\mathbb{E}|\widehat{\widehat{\nu}_X}(k) - \widehat{\bar{\nu}}(k)|^2}{4\pi^2|k|^2} \;\le\; \frac{b_K S}{4\pi^2} \sum_{k\neq 0} \frac{1}{|k|^2}\Big(1 \wedge \Big(\frac{R}{|k|}\Big)^{\kappa}\Big),$$

$$R := \frac{C_Q}{h_\star} = \frac{1}{D\theta_{\mathrm{sm}}}.$$

Split at $|k| = R$. The low block is $\sum_{0<|k|\le R} |k|^{-2} \le c_d R^{d-2}$ for $d \ge 3$; the high block is $R^{\kappa}\sum_{|k|>R} |k|^{-2-\kappa} \le c_d R^{\kappa} R^{d-2-\kappa} = c_d R^{d-2}$, the sum converging because $2+\kappa > d$. Hence $\mathbb{E}\|\cdot\|_{\dot{H}^{-1}}^2 \le c_{d,K} S \theta_\star^{2-d}$ with $\theta_\star = \theta_{\mathrm{sm}}$, and Jensen gives (47). □

Remark C.9 ($\kappa - (d-2)$ is a capacity threshold). The convergence of the high block is not an artefact of the estimate. $\|\sigma\|_{\dot{H}^{-1}}^2$ is the Newtonian energy $c_d \iint |x-y|^{2-d} d\sigma\, d\sigma$, and a measure carried by a κ-dimensional set has finite energy iff $\kappa > d-2$, the classical capacity criterion, since $\int_0^1 r^{2-d} r^{\kappa-1} dr < \infty$ exactly then. So the sign of $\kappa - (d-2)$ in Proposition B.4 is not a chaining exponent that a sharper argument might improve: it decides whether the linearised transport energy of the synthetic measure is finite at all, and the convergence of the dyadic sum as $\theta_L \to 0$ in Section B Step 4 is that condition.

Below the threshold (D2) is worse than unproven. The support of $\widehat{\nu}_X$ is a union of n_π sets of dimension κ, and a smooth $\bar{\nu}$ has mass that no coupling can move cheaply onto such a thin set.

Proposition C.10 (A covering lower bound, and what it costs)**.** *Let Σ be the union of the* distinct *simplices drawn, n_Σ of them, each of dimension at most κ and diameter at most $h_\star$, and let $\bar{\nu}$ have density at most $f_{\max}$. Then $\bar{\nu}(\Sigma^{\oplus r}) \le c_d f_{\max}\, n_\Sigma\, h_\star^{\kappa} r^{d-\kappa}$ for $r \le h_\star$, and therefore, almost surely,*

$$W_1(\widehat{\nu}_X, \bar{\nu}) \;\ge\; \tfrac{1}{2} r_\star, \qquad r_\star := \min\Big\{h_\star,\; \big(2c_d f_{\max}\, n_\Sigma\, h_\star^{\kappa}\big)^{-1/(d-\kappa)}\Big\}. \tag{48}$$

Consequently, writing $G := \theta_\star/\theta_{\mathrm{samp}}$ for the ratio of scales in (31) and taking balanced reuse $S = 1/n$: at $n_\Sigma \asymp n$, $r_\star$ matches (37), so Proposition B.4(iii) is sharp; and a sub-critical bound with the factor $G^{-\kappa/2}$ that (D2) would give forces

$$\frac{n_\Sigma}{n} \;\gtrsim\; G^{\,\kappa(d-\kappa-2)/2}. \tag{49}$$

Proof. Any coupling must move $\bar{\nu}$-mass at y to a point of $\Sigma = \operatorname{supp}\widehat{\nu}_X$, so $W_1(\widehat{\nu}_X, \bar{\nu}) \ge \int \operatorname{dist}(y,\Sigma)\, d\bar{\nu}(y)$. A κ-dimensional convex set of diameter $h_\star$ has κ-volume $\le c h_\star^{\kappa}$, so its r-neighbourhood has volume $\le c_d(h_\star^{\kappa} r^{d-\kappa} + r^d) \le 2c_d h_\star^{\kappa} r^{d-\kappa}$ for $r \le h_\star$; summing over α and multiplying by $f_{\max}$ gives the mass bound. At $r = r_\star$ that mass is at most $\frac{1}{2}$, so $\bar{\nu}(\{\operatorname{dist}(\cdot,\Sigma) \ge r_\star\}) \ge \frac{1}{2}$ and (48) follows. With $n_\Sigma = n = \theta_{\mathrm{samp}}^{-d}$ and $h_\star \asymp \theta_\star$, $r_\star \asymp \theta_{\mathrm{samp}}^{d/(d-\kappa)} \theta_\star^{-\kappa/(d-\kappa)}$, the right-hand side of (37). For (49), the (D2)-based bound is $B \asymp \theta_{\mathrm{samp}}^{(\kappa+2)/2} \theta_\star^{-\kappa/2}$; imposing $r_\star \le B$ and substituting $h_\star \asymp \theta_\star$, $n = \theta_{\mathrm{samp}}^{-d}$, $\theta_\star = G\theta_{\mathrm{samp}}$ makes every power of θ_{samp} cancel and leaves (49). □

The exponent in (49) is negative exactly when $\kappa > d-2$, where $n_\pi \ge n$ already suffices and balanced reuse supplies it; it vanishes at criticality; and it is positive only in the sub-critical regime, where it is a genuine constraint that the overlap factor cannot see. For mixup, $\kappa = 1$, it reads $n_\Sigma \gtrsim nG^{(d-3)/2}$, and since sampling is with replacement it is the number of *distinct* pairs that must be that large. The two generators of Section 5.3 answer differently. For k-NN SMOTE $h_\star \asymp D(k/n)^{1/d}$ gives $G = k^{1/d}/C_Q$, with n cancelling as in (86), and the condition asks for $nk^{\kappa(d-\kappa-2)/(2d)}$ distinct tuples out of the $\asymp nk^{\kappa}$ available: richness is available there, at $n_\pi \gg n$ draws. For global mixup $h_\star = \Theta(D)$ gives $G \asymp n^{1/d}$ and the condition becomes $n_\Sigma \gtrsim n^{1+(d-3)/(2d)}$, which is more distinct pairs than real points, by a factor $n^{(d-3)/(2d)}$ that grows from $n^{1/8}$ at $d=4$ towards $\sqrt{n}$; available, since $\binom{n}{2}$ pairs exist, but not automatic. This is the one place in the paper where n_π enters other than through w, and it does not contradict Remark A.5: more combinations still cannot beat the n-dependent scale, they can only realise a resolution already permitted by it.

Remark C.11 (Relation to known smoothing results). Theorem B.3 instantiates a phenomenon established for other smoothing mechanisms. Goldfeld et al. (2020) show that convolving an empirical measure with a Gaussian of fixed width makes W_1 to the correspondingly smoothed population distribution converge at the parametric rate $e^{O(d)}n^{-1/2}$ instead of $n^{-1/d}$, the same trade of a dimension-free exponent against a constant exponential in d that (28) shows through $\Phi_d(\theta_{\mathrm{sm}}) \asymp (C_Q D/h_\star)^{d/2-1}$ when $\theta_\star = \theta_{\mathrm{sm}}$ and $d \geq 3$. On the transport side, the rigorous form of "both measures are smooth at scale $h_\star$, so the local cost is governed by the density discrepancy rather than the cell diameter" is the quantitative linearisation of Goldman et al. (2021), and the heat-semigroup regularisation of Ambrosio et al. (2019) is the same device in the matching problem. Those are the two natural routes to (D2) at low arity, and Proposition C.8 is what they give: Peyre's inequality is the crudest form of the linearisation, and the object it linearises around, the $\dot{H}^{-1}$ energy, is what the Monge–Ampère argument expands and what the heat flow is introduced to render admissible. They stop at $\kappa = d - 2$ for a structural reason rather than a technical one, since that energy is infinite below it (Remark C.9).

Regularisation does not rescue (D2) in the sub-critical case either, and the accounting is one line. Mollifying at scale $\sqrt{\tau}$ makes both measures absolutely continuous, but for the resulting density bound to replace the finest-cell term at θ_{samp} the mollifier must at least fill that cell transversally, $\sqrt{\tau} \gtrsim D\theta_{\mathrm{samp}}$, while the displacement it costs is $W_2(\mu, P_\tau\mu) \leq \sqrt{2\tau d}$ for every μ, already of the baseline order $D\theta_{\mathrm{samp}}$, which is what the refinement is trying to beat. The heat flow pays exactly what it is meant to save. At $n_\Sigma \asymp n$ the covering bound is the whole story: (37) is its matching upper bound, obtained by keeping the finest-cell term instead of regularising it away. A direct simulation of the covering distance reproduces the common exponent $1/(d-\kappa)$, with fitted values $0.48, 0.31, 0.23$ against $0.50, 0.33, 0.25$ for $d = 3, 4, 5$ at $\kappa = 1$. What remains open is $n_\Sigma \gg n$: whether richer generators reach the factor $G^{-\kappa/2}$, a random-matching problem in which the pieces are κ-dimensional rather than points, for which Ambrosio et al. (2019) is the natural template and no result appears to exist. Until it is settled, every sub-critical exponent in the paper is exact at $n_\Sigma \asymp n$ and only a lower bound on the true exponent above it, since (37) does not depend on n_π while (48) falls with n_Σ. For global mixup the proved fluctuation is $n^{-1/(d-1)}$; with all $\binom{n}{2}$ pairs the covering bound drops to $n^{-2/(d-1)}$, and no statistic can fluctuate below $n^{-1/2}$ since $S \geq 1/n$, so the true exponent lies in $[\frac{1}{d-1}, \min\{\frac{2}{d-1}, \frac{1}{2}\}]$. Here n_Σ counts the simplices carried by the synthetic distribution, not synthetic points. The generator of Theorem A integrates over all $\binom{k}{K-1}$ partner sets of each base, so $n_\Sigma = n_B\binom{k}{K-1} \gg n$ and (48) at this n_Σ lies below (37) by a factor of order $k^{-\kappa/(d-\kappa)}$; the bound of Lemma E.3 also averages $\Phi_x^\circ(Q)^2$ over partner sets, which for a star of k segments gives away up to a factor k at fine scales. Lemma E.9 recovers this factor for the bases, and Proposition E.10 gives the exponents that follow if the pool obeys the same two-point bound (Assumption 3). For that generator, as for mixup drawing fresh pairs in every batch, the sub-critical exponents of Theorem A are therefore lower bounds, which are loose if Assumption 3 holds, and Theorem B caps the true exponent at $2/(d+2)$. A sample of synthetic points is a different matter: N of them cannot beat $N^{-1/d}$ whatever distribution they come from (Corollary 3.1), so a SMOTE run that draws $N \asymp n$ points cannot beat $n^{-1/d}$ in W_1. The statements that tuned mixing beats $n^{-1/d}$ rest on the upper bound alone and are unaffected.

D THE MIXING BIAS

The locality scale controls not only the resolution scale $\theta_\star$ but the mixing bias itself, by an argument that needs nothing beyond the coupling already used in Section A.4.

Proposition D.1 (Mixing bias of a local generator)**.** *Suppose $\delta_\alpha(X) \leq h_\star$ almost surely for every $\alpha \in A$. Then*

$$W_1(\widehat{\nu}_X, \mu_w) \;\leq\; h_\star \qquad \textit{almost surely,} \tag{50}$$

so in particular $\eta_{\mathrm{int},1} \leq h_\star$. If moreover the weights w do not depend on X, then $\mathbb{E}\mu_w = \mu_0$ and $\beta_1 = W_1(\bar{\nu}, \mu_0) \leq h_\star$. For a data-adaptive index set the same argument gives $\beta_1 \leq h_\star + W_1(\mathbb{E}\mu_w, \mu_0)$, the second term being the selection bias of the generator.

Proof. Draw $\alpha \sim \mathrm{Unif}(A)$, $u^\alpha \sim \mathrm{Dir}(1, \dots, 1)$ and $j \sim \mathrm{Unif}\{1, \dots, K\}$ independently. The pair $(Y_\alpha(X, u^\alpha), X_{\alpha_j})$ has first marginal $\widehat{\nu}_X$ and second marginal μ_w, because $\mathbb{P}(X_{\alpha_j} = X_i) =$

$\frac{1}{mK}\#\{(\alpha,j) : \alpha_j = i\} = w_i$; it is therefore a coupling. Since $Y_\alpha - X_{\alpha_j} = \sum_l u_l^\alpha (X_{\alpha_l} - X_{\alpha_j})$ with $\sum_l u_l^\alpha = 1$,

$$\|Y_\alpha - X_{\alpha_j}\|_2 \le \sum_l u_l^\alpha \|X_{\alpha_l} - X_{\alpha_j}\|_2 \le \max_{l,l'} \|X_{\alpha_l} - X_{\alpha_{l'}}\|_2 = \delta_\alpha(X) \le h_\star,$$

which gives (50). For the second claim, W_1 is jointly convex in its two arguments, so Jensen's inequality gives $W_1(\mathbb{E}\widehat{\nu}_X, \mathbb{E}\mu_w) \le \mathbb{E}W_1(\widehat{\nu}_X, \mu_w) \le h_\star$, and $\mathbb{E}\mu_w = \mu_0$ identifies the left-hand side as β_1. The third claim is the triangle inequality. □

Proposition D.1 is first order in $h_\star$ and, for a smooth population distribution, not tight: convex mixing with exchangeable weights is a *symmetric* smoothing operation, so the first-order displacement cancels and only curvature survives.

Lemma D.2 (Moments and flux of local mixing)**.** *In the setting of Proposition 4.1 with $\alpha = 1$, write $h_\star$ for the radius h of the partner ball and put $\Delta := Y - Z_1$ and*

$$\begin{aligned} b(z) &:= \mathbb{E}[\Delta \mid Z_1 = z] = \frac{K-1}{K}\,\frac{h_\star^2}{d+2}\,\nabla \log f(z) + O(h_\star^3), \\ M(z) &:= \mathbb{E}[\Delta\Delta^\top \mid Z_1 = z] = \frac{2(K-1)}{K(K+1)}\,\frac{h_\star^2}{d+2}\,I_d + O(h_\star^3). \end{aligned} \tag{51}$$

Then $\bar{\nu}_h - \mu_0 = -\nabla \cdot V_0 + R$ with $\|R\|_{L^1} = o(h_\star^2)$ and

$$V_0 \;=\; fb - \tfrac{1}{2}\nabla \cdot (fM) \;=\; \frac{K-1}{(d+2)(K+1)}\,h_\star^2\,\nabla f + O(h_\star^3). \tag{52}$$

By the Beckmann formulation of W_1 the lemma gives (11) at $\alpha = 1$, where $C_\partial \le c_{d,K} f_{\max} \operatorname{Per}(\Omega)$ is the contribution of the layer of width $h_\star$ along $\partial\Omega$, and $C_\partial = 0$ when Ω has no boundary (a torus) or f vanishes on $\partial\Omega$. The boundary term is of the same order as the interior one, so the explicit constant is the whole constant only in those two cases. In dimension $d = 1$ on the circle the Beckmann problem is solved explicitly by the antiderivative, so $\beta_1 = \frac{K-1}{3(K+1)} h_\star^2 \min_a \|f' - a\|_{L^1} + o(h_\star^2)$, the constant a being the flux left free by periodicity. It vanishes when f' has median zero, for instance when f is symmetric, and then $\beta_1 = \frac{K-1}{3(K+1)} h_\star^2 \|f'\|_{L^1} + o(h_\star^2)$, which is (12) at $\alpha = 1$. Every $o(h_\star^2)$ here is $O(h_\star^{2+\gamma})$ when $D^2 f$ is γ-Hölder, and $O(h_\star^3)$ when $f \in C^3$. Here $h_\star$ is the radius of the partner ball, as in the numerical checks of Sections G and I; the spread of Definition B.1 is then at most $2h_\star$ ($h_\star$ for $K = 2$), and where Theorem A′ writes $h_\star = 2r_{\max}$ for the spread, (11) applies with $r_{\max}$ in place of $h_\star$.

Proof of Lemma D.2, and of Proposition 4.1 at $\alpha = 1$. Write $T_j := Z_j - Z_1$ for $j \ge 2$, so $\Delta = \sum_{j\ge 2} U_j T_j$. Since $\int_{B(0,h)} t\,t^\top dt = \frac{h^2}{d+2}\mathrm{vol}(B(0,h)) I_d$ and $\int_{B(0,h)} t\,dt = 0$, a Taylor expansion of f inside $B(z, h_\star)$ gives $\mathbb{E}[T_j \mid Z_1 = z] = \frac{h_\star^2}{d+2}\nabla \log f(z) + O(h_\star^3)$ and $\mathbb{E}[T_j T_j^\top \mid Z_1 = z] = \frac{h_\star^2}{d+2} I_d + O(h_\star^3)$; the first identity is where the $O(h_\star)$ term cancels, by symmetry of the ball. Using $\mathbb{E}U_j = 1/K$ and $\mathbb{E}U_j^2 = 2/(K(K+1))$ for the flat Dirichlet distribution, and discarding the cross terms $\mathbb{E}[U_j U_l]\mathbb{E}[T_j]\mathbb{E}[T_l] = O(h_\star^4)$, gives (51). The Kramers–Moyal expansion of the transition kernel then yields $\bar{\nu} - \mu_0 = -\nabla \cdot (fb) + \frac{1}{2}\nabla \cdot \big(\nabla \cdot (fM)\big) + R$, where R is the remainder of the second-order Taylor expansion, in the base point, of $z \mapsto f(z)p(\delta \mid z)$, with $p(\cdot \mid z)$ the distribution of Δ given $Z_1 = z$: its second derivatives involve f up to $D^2 f$ on $B(z, h_\star)$ and inherit the modulus of continuity $\omega_{D^2 f}$ of $D^2 f$ on $\bar{\Omega}$, so $|R| \lesssim h_\star^2\, \omega_{D^2 f}(h_\star)$ and $\|R\|_{L^1} = o(h_\star^2)$, which is (52); substituting (51) and $\nabla \cdot (fM) = \frac{2(K-1)h_\star^2}{K(K+1)(d+2)}\nabla f$ gives the closed form, the two contributions combining as $\frac{K-1}{K(d+2)}\big(1 - \frac{1}{K+1}\big) = \frac{K-1}{(d+2)(K+1)}$. Finally $W_1(\bar{\nu}, \mu_0) = \min\{\int |V| : \nabla \cdot V = \mu_0 - \bar{\nu}\}$, so any admissible field bounds β_1; taking $V = V_0$ and absorbing R (a mean-zero L^1 function on a bounded domain, transportable at cost $O(\operatorname{diam}\Omega\, \|R\|_{L^1})$) gives (11) away from $\partial\Omega$. Within $h_\star$ of $\partial\Omega$ the ball is truncated and b, $h_\star^{-1}\nabla \cdot (fM)$ are $O(h_\star)$ rather than $O(h_\star^2)$; that layer has volume $O(\operatorname{Per}(\Omega) h_\star)$, so it adds $C_\partial h_\star^2$ to $\int |V_0|$, the same order as the interior term. If f vanishes on $\partial\Omega$ the layer carries mass $O(h_\star^2)$ and the addition is $O(h_\star^3)$. The boundary term is genuine: for μ_0 uniform on $[0,1]$ and $K = 2$ the interior term vanishes, while a deterministic quadrature gives $\beta_1/h_\star^2 = 0.1389$ for $h_\star \in [0.025, 0.2]$, i.e. $\beta_1 \approx \frac{5}{36} h_\star^2$. □

Remark D.3 ($\mathrm{Dir}(\alpha)$ weights, and where the bias comes from). With $U \sim \mathrm{Dir}(\alpha,\dots,\alpha)$ only the second moment of the weights changes: $\mathbb{E}U_j = 1/K$ and $\mathbb{E}U_j^2 = (\alpha+1)/(K(K\alpha+1))$. The drift b of (51) is unchanged, M is multiplied by $(\alpha+1)(K+1)/(2(K\alpha+1))$, and the proof of Proposition 4.1 gives $V_0 = c_K(\alpha)h_\star^2\nabla f + O(h_\star^3)$ with

$$c_K(\alpha) \;=\; \frac{(K-1)\big((2K-1)\alpha+1\big)}{2K(K\alpha+1)(d+2)},$$

in place of $(K-1)/((K+1)(d+2))$, its value at $\alpha=1$, which gives (11) and (12) for every $\alpha>0$. It increases with α, from $c_K(0) = (K-1)/(2K(d+2))$ to $(K-1)(2K-1)/(2K^2(d+2))$. The limit $\alpha\to 0$ is the generator that outputs a uniformly chosen vertex of the tuple, with no interpolation, and its bias is that of choosing a neighbour in a density gradient alone; at $\alpha=1$ this is already $c_K(0)/c_K(1) = (K+1)/(2K)$ of the whole, three quarters for pairs, and small α therefore removes at most the fraction $(K-1)/(2K)$ of the bias. These are the numbers for a fixed radius. The k-NN radius adapts to the density, $\nabla\rho_q^2 = -\frac{2}{d}\rho_q^2\nabla\log f + O(\rho_q^3)$ (Lemma E.5), the divergence in V_0 also hits it, and the same computation gives $V_0 = c_K^{\mathrm{nn}}(\alpha)\rho_q^2\nabla f + o(\rho_q^2)$ with

$$c_K^{\mathrm{nn}}(\alpha) \;=\; \frac{K-1}{K(d+2)}\Big(1-\frac{(\alpha+1)(d-2)}{2d(K\alpha+1)}\Big),$$

the constant of Lemma E.5 at $\alpha=1$. The vertex choice then carries $c_K^{\mathrm{nn}}(0)/c_K^{\mathrm{nn}}(1) = (d+2)(K+1)/(2(dK+2))$ of the bias, 0.94 for pairs at $d=3$ and 0.90 at $d=4$, and three quarters only as $d\to\infty$: with a density-adapted radius, neighbour selection is almost the whole bias, and for $d\ge 3$ small α removes at most the fraction $(d-2)(K-1)/(2(dK+2))$ of it. Measured at $d=4$ on the cosine torus with $\rho(z) = hf(z)^{-1/4}$, which makes $\nabla(f\rho^2) = (1-\frac{2}{d})\rho^2\nabla f$ exact, and with the same bases and partners for every weight distribution, the ratio is 0.906 ± 0.004 for pairs and 0.860 ± 0.004 at $K=3$, against 0.900 and 0.857, and $c_K^{\mathrm{nn}}(\alpha)$ agrees with its closed form within 1.1% at $\alpha = 0, 0.3, 1, 3$. Small α costs smoothing: for pairs, with U,U' i.i.d. $\mathrm{Beta}(\alpha,\alpha)$, $\mathbb{P}(|U-U'|\le\delta) \asymp \delta^{\min\{2\alpha,1\}}$, with a logarithm at $\alpha=\frac12$, because both weights fall within δ of the same end with probability $\asymp\delta^{2\alpha}$ and $U-U'$ has a bounded density for $\alpha>\frac12$. The proof of Proposition C.3 then gives Assumption 2 with $\kappa = \min\{2\alpha,1\}$, and so does that of Corollary A.13 for global pairs, which therefore estimate their own population distribution at $n^{-1/(d-\min\{2\alpha,1\})}$ for $d\ge 3$: mixup at $\alpha<\frac12$ trades smoothing for at most a quarter of the bias, and α is a second lever next to k. Measured with the estimator behind Table 3 (fixed radius, $d=4$), c_K is 0.0468, 0.0518, 0.0591 at $K=2$, $\alpha = 0.2, 0.5, 2$, against 0.0476, 0.0521, 0.0583, and 0.0727, 0.0900 at $K=3$, $\alpha = 0.3, 3$, against 0.0731, 0.0889; the vertex choice gives 0.0421 and 0.0559 against 0.0417 and 0.0556; the fitted pair exponents are 0.30, 0.50, 0.69, 0.89, 1.00 at $\alpha = 0.15, 0.25, 0.35, 0.5, 0.8$. Two things stay open. Theorem A is proved for $\mathrm{Dir}(1,\dots,1)$: its pool term (Lemma E.4) uses the one-point bound (57), whose exponent for $\mathrm{Beta}(\alpha,\alpha)$ pairs is $\min\{\alpha,1\}$, not $\min\{2\alpha,1\}$. And for $K\ge 3$ the small-ball exponent under $\mathrm{Dir}(\alpha)$ remains to be computed.

Remark D.4 (The bias is a genuine h^2, not an upper bound). Because the $d=1$ formula of Proposition 4.1 is an equality, β_1 is bounded *below* on the circle by $\frac{K-1}{3(K+1)}h^2\min_a\|f'-a\|_{L^1}$, with h the partner radius of Proposition 4.1, which is (12) when f' has median zero: the mixing bias vanishes only when $\nabla f\equiv 0$, and is $\Theta(h^2)$ otherwise. That equality is a $d=1$ statement, but the lower bound is not. In any dimension the expansion of Proposition 4.1 gives $\int g\,d(\bar\nu-\mu_0) = c_Kh^2\int\nabla g\cdot\nabla f + o(h^2)$ for 1-Lipschitz g on a domain without boundary, so

$$\beta_1 \;\ge\; c_Kh^2 \sup_{\mathrm{Lip}(g)\le 1}\int \nabla g\cdot\nabla f \;+\; o(h^2), \tag{53}$$

and the supremum vanishes only if $\Delta f = 0$, hence only for constant f on the torus. So $\beta_1 = \Theta(h^2)$ holds generally, with the dual constant of (53); it reduces to $\min_a\|f'-a\|_{L^1}$ in $d=1$, and to $\|f'\|_{L^1}$ when in addition f' has median zero. A skewed density on the circle separates the two, 3.9378 against 3.9893 in a numerical check. On a domain with boundary it is $\Theta(h^2)$ even for a uniform distribution: the collar of width h is depleted by mass $\Theta(h)$, which has to travel $\Theta(h)$, which is C_∂. On the circle with $f = 1 + a\cos 2\pi x$ and $K=2$, (12) predicts $\beta_1/h^2 = 4a/9$. A deterministic quadrature of the distribution of Y reproduces this constant to within 0.5% at $h=0.02$ for $a\in\{0.2,0.3,0.6,0.9\}$; above that the $O(h)$ remainder of (11) shows, the ratio to the closed form being 0.97–0.99 at $h=0.05$ and 0.82–0.89 at $h=0.14$, and the fitted log–log slope over $h\in[0.02,0.14]$ is 1.91–1.95. Monte Carlo agrees (0.871 at $h=0.14$ and 0.987 at $h=0.04$ for $a=0.6$).

Mixing-bias constant across arities. Circle $[0,1)$, $f = 1 + a\cos 2\pi x$ with $a = 0.6$ (so $\|f'\|_{L^1} = 4a$), base point $Z_1 \sim f$, partners drawn from f restricted to $[Z_1 - h, Z_1 + h]$ by rejection, $U \sim \mathrm{Dir}(1,\dots,1)$. We estimate $F_Y - F$ by $\frac{1}{N}\sum_k[\mathbf{1}\{Y_k \le y\} - \mathbf{1}\{Z_{1,k} \le y\}]$ on a 4096-cell grid: the Z_1-noise cancels and the variance is $O(h/N)$, which is what makes an $O(h^2)$ signal measurable, and $\beta_1 = \min_c \int |F_Y - F - c|$. At fixed h the ratio to the closed form is nearly the same for every arity (Table 2), so the relative $O(h)$ remainder of (11) hardly depends on K; at $K = 2$ its decay as $h \downarrow 0$ is quantified in Remark D.4.

Table 2: Mixing-bias constant on the circle, measured/predicted across arities at fixed partner radius h; $N = 8 \times 10^6$ draws per cell.

ratio measured/predicted	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$h = 0.16$	0.837	0.844	0.846	0.852
$h = 0.12$	0.909	0.906	0.909	0.914
$h = 0.09$	0.941	0.948	0.951	0.946

A kernel smoother at the same rate. The rate of Theorem A for $K \ge d - 1$ is that of a tuned Gaussian smoother.

Lemma D.5 (Tuned Gaussian smoothing)**.** *Let $d \ge 3$, let μ_0 on $\mathbb{T}^d$ have a density $f \in C^1$, and let $\mathcal{N}_\sigma$ be the Gaussian distribution of covariance $\sigma^2 I_d$, wrapped on $\mathbb{T}^d$. Then*

(i) $W_1(\mu_0 * \mathcal{N}_\sigma, \mu_0) \le \frac{1}{2}\sigma^2\|\nabla f\|_{L^1}$;

(ii) $\mathbb{E}W_1(\hat\mu_n * \mathcal{N}_\sigma, \mu_0 * \mathcal{N}_\sigma) \le c_d\, n^{-1/2}\sigma^{1-d/2}$ *for* $n^{-1/d} \le \sigma \le 1$, *with c_d depending on d only;*

so at $\sigma = n^{-1/(d+2)}$, $\mathbb{E}W_1(\hat\mu_n * \mathcal{N}_\sigma, \mu_0) \le (\frac{1}{2}\|\nabla f\|_{L^1} + c_d)\, n^{-2/(d+2)}$.

Proof. (i) With P_t the heat semigroup, $\mu_0 * \mathcal{N}_\sigma = P_{\sigma^2}\mu_0$ and $\partial_t P_t f = \frac{1}{2}\Delta P_t f$. For 1-Lipschitz ϕ, $\int \phi\, d(P_{\sigma^2}\mu_0 - \mu_0) = -\frac{1}{2}\int_0^{\sigma^2}\int \nabla\phi \cdot \nabla P_t f\, dt \le \frac{1}{2}\int_0^{\sigma^2}\|P_t\nabla f\|_{L^1} dt \le \frac{1}{2}\sigma^2\|\nabla f\|_{L^1}$. (ii) Put $\hat\nu := \hat\mu_n * \mathcal{N}_\sigma$, so $\mathbb{E}\hat\nu = \mu_0 * \mathcal{N}_\sigma$ and, for a cell Q of side s, $\mathrm{Var}\,\hat\nu(Q) \le n^{-1}\mathbb{E}[\mathcal{N}_\sigma(Q - X)^2]$. Summed over a dyadic partition, $\sum_Q \mathcal{N}_\sigma(Q - x)^2$ is the probability that $x + G$ and $x + G'$, with G, G' independent of distribution $\mathcal{N}_\sigma$, share a cell, at most $\mathbb{P}(\|G - G'\| \le \sqrt{d}s) \le C_d(s/\sigma \wedge 1)^d$, as the wrapped distribution of $G - G'$ has density at most $C_d\sigma^{-d}$ for $\sigma \le 1$. This is (D1) with $S = 1/n$, $b_K = 1$, exponent d and σ in place of $h_\star$, and Steps 1–2 and the case $\kappa > d - 2$ of Step 4 in Section B, which use (D1) only, give (ii) with $\theta_{\mathrm{sm}} = \sigma/C_Q$ above the sampling scale $n^{-1/d}$ up to a constant. The last claim adds (i) and (ii). □

On a bounded domain, (i) holds for f extended by zero with $\|\nabla f\|_{L^1}$ replaced by the total variation of that extension, which adds at most $f_{\max}\mathrm{Per}(\Omega)$: mass leaking across $\partial\Omega$ costs $O(\sigma^2)$ in W_1, the order of the collar term $C_\partial h_\star^2$ of Proposition 4.1. At a boundary the two smoothers therefore differ in the constant of the bias, not in its order.

Remark D.6 (A radius that grows with the density). The computation in the proof of Lemma E.5 applies to any partner radius $\rho(z)$ that is a C^1 function of the base. At flat weights the drift is $b = \frac{K-1}{K}\frac{\rho^2}{d+2}\nabla\log f$ and the second moment is $M = \frac{2(K-1)}{K(K+1)}\frac{\rho^2}{d+2}I_d$, to leading order, so

$$V_0 = fb - \tfrac{1}{2}\nabla\cdot(fM) = c_K(1)\Big(\rho^2\nabla f - \frac{f}{K}\nabla\rho^2\Big) + o(\rho^2).$$

For the k-NN radius, $\rho \propto f^{-1/d}$, this is $c_K(1)(1 + \frac{2}{dK})\rho^2\nabla f$, the constant of Lemma E.5. For $\rho = hf^{K/2}$ the two terms cancel and the interior bias is $o(h^2)$, although every weight is non-negative. This is the mechanism of Abramson's square-root law (Abramson, 1982), where a bandwidth that adapts to the density removes the second-order bias of a positive kernel. On the circle with $f = 1 + 0.6\cos 2\pi x$ and $K = 2$, a deterministic quadrature of the distribution of Y gives $\beta_1/(\frac{4a}{9}h^2) = 0.98$ and 1.00 at $h = 0.05$ and 0.025 for a fixed radius, 2.92 and 3.07 for $\rho = hf^{-1}$ (limit 3.125 by the display), and 0.013 and 0.003 for $\rho = hf$, whose local log–log slope in h is 3.9. Such a rule needs f, or a pilot estimate of it. It falls outside Theorem B, whose bias hypothesis Lemma E.7 proves for radii with $f\nabla f \cdot \nabla\rho^2 \le 0$, and whether a data-driven version beats $n^{-2/(d+2)}$ is open.

E PROOFS OF THEOREMS A AND B

We put $r_{\max} := (4k/(c_0 n_A))^{1/d}$, which exceeds every k-NN radius with high probability.

Theorem A′ (Convex augmentation in W_1)**.** *Let $\widehat{\nu}_X$ be the synthetic distribution of one of the following generators:*

(G1) *the reuse generator of Definition A.1 on i.i.d. data with* $\operatorname{diam}(\operatorname{supp}\mu_0) \le D$ *and pairwise distinct indices in every tuple;*

(G2) *k-NN SMOTE of Definition 2.1 with $\alpha = 1$ on i.i.d. data, under Assumption 1 with $K \ge 2$, $k \ge K-1$ and $r_{\max} \le D$.*

Put $\bar{\nu} := \mathbb{E}\widehat{\nu}_X$, $\beta_1 := W_1(\bar{\nu}, \mu_0)$ and $\kappa := \min(K-1, d-1)$; let S be the overlap and $h_\star$ the scale of Definition B.1 in (G1), and $S := 1/n_A$, $h_\star := 2r_{\max}$ in (G2); and let $\theta_\star$, $R_{d,\kappa}$ be as in Definition B.2. Then:

(i) fluctuation: $\mathbb{E}\, W_1(\widehat{\nu}_X, \bar{\nu}) \le C_{d,K} D\, \psi_d(S)$, *and, under Assumption 2 in (G1) and unconditionally in (G2),*

$$\mathbb{E}\, W_1(\widehat{\nu}_X, \bar{\nu}) \;\le\; c_{d,K} D\big(\sqrt{S}\,\Phi_d(\theta_\star) + R_{d,\kappa}(\theta_\star)\big); \tag{54}$$

in (G2) both bounds carry an additional $2De^{-k/2}$, and constants depending on c_f;

(ii) deviation: *for every $\delta \in (0,1)$, with probability at least $1-\delta-p_n$,*

$$\big|W_1(\widehat{\nu}_X, \mu_0) - \mathbb{E}W_1(\widehat{\nu}_X, \mu_0)\big| \;\le\; \Delta_\delta, \tag{55}$$

where $\Delta_\delta = D\sqrt{S\log(2/\delta)/2}$ and $p_n = 0$ in (G1), and $\Delta_\delta = 21(K-1)\, c_f\, D\sqrt{\log(2/\delta)/n}$ and $p_n = ne^{-k/2}$ in (G2);

(iii) tuned rate: *in (G2) with $d \ge 3$, $\beta_1 = O(h_\star^2)$, and with $\bar{\kappa} := \min(K-1, d-2)$ and k as in (2), there are C and n_0, depending only on d, K, Ω and f, such that $\mathbb{E}\, W_1(\widehat{\nu}_X, \mu_0) \le C\, n^{-2/(2d-\bar{\kappa})}\, \ell_n$ for $n \ge n_0$, where $\ell_n = \log n$ if $K = d-1$ and $\ell_n = 1$ otherwise.*

In (G1) the tuples ignore the data, so $h_\star \asymp D$ and $\beta_1 = W_1(\mathcal{S}_K(\mu_0), \mu_0)$ does not vanish; (i) is Theorems A.7 and B.3 and measures how well $\widehat{\nu}_X$ estimates its own population distribution (Corollary A.13). In (G2) locality makes $\beta_1 = O(h_\star^2)$ (Lemma E.5) and tuning possible, and the deviation (ii) comes from bounded differences (Lemma E.6). For $K > d$ the small-ball estimate behind the fluctuation is applied to a d-point sub-tuple, which keeps the exponent $d-1 > d-2$: arity above d changes constants, not the rate, and $K = d$ already reaches $n^{-2/(d+2)}$ without the logarithm. n_0 is set by $k > C_Q^d$, the form the threshold (30) takes here, a worst case (a gain is observed from $n = 90$, Section G), and by the range of the bias expansion; the truncation term $2De^{-k/2}$ falls below the rate once $k \gtrsim \log n$.

Remark E.1 (The ε-ball rule). If the partners are drawn among the pool points within distance ε of X_b instead, copies of X_b filling in when there are fewer than $K-1$, Theorem A′ holds for (G2) verbatim with $r_{\max} := \varepsilon$, $k := c_0 n_A \varepsilon^d \ge 2K-2$ (the least expected number of pool points in the ball), $e^{-k/8}$ in place of $e^{-k/2}$, and $\varepsilon = \varepsilon_0 n^{-1/(2d-\bar{\kappa})}$ in (iii), C and n_0 then depending on ε_0 too. The only other change is the bias, which is Proposition 4.1 itself, without the factor $1 + 2/(dK)$ (proved below).

Proof of Theorem A. Under (2), $k \ge K-1$ and $r_{\max} = (4k/(c_0 n_A))^{1/d} \le D$ once $n \ge n_0$, so the generator of Theorem A is (G2) of Theorem A′; its part (i) is Theorem A′(iii), and its part (ii) is Theorem A′(ii) at this k, where $p_n = ne^{-k/2}$. □

On the torus. With Assumption 1 read on $\mathbb{T}^d$ as in Section F, the proof of Theorem A′ below goes through verbatim: the dyadic cubes of $[0,1)^d$ partition $\mathbb{T}^d$ and the periodic distance is at most the chart distance, Lemma E.2 is local and holds once $r_{\max} \le \frac{1}{8}$ (as in the proof of Theorem F.1), and the bias of Lemma E.5 has $C'_\partial = 0$. So Theorem A holds on $\mathbb{T}^d$ at every $K \ge 2$; for $K \ge d$ it also follows from Theorem F.1(iii), since $W_1 \le W_2$. This is the setting of the lower bound Theorem B.

The reuse model (G1). Part (i) is Theorem A.7, whose constant carries $\sqrt{b_K} \le K$, and Theorem B.3. For (ii), the map $X \mapsto W_1(\widehat{\nu}_X, \mu_0)$ is coordinatewise Lipschitz with constants w_i, by Step 8 of Section A.4 with μ_0 in place of $\bar{\nu}$, so it has bounded differences $w_i D$ and McDiarmid's inequality gives $\mathbb{P}(|W_1(\widehat{\nu}_X, \mu_0) - \mathbb{E}W_1(\widehat{\nu}_X, \mu_0)| \ge t) \le 2e^{-2t^2/(D^2 S)}$; under log-Sobolev inequalities Steps 3–5 of Section A.3 give $2e^{-t^2/(2V_1)}$ instead. Part (iii) does not concern (G1). The rest of this appendix treats (G2), k-NN SMOTE, and alongside it the ε-ball rule of Remark E.1: the two rules. It ends with the proof of Theorem F.1 for (G2).

Setting for the SMOTE rules. Throughout, the data are i.i.d. μ_0, Assumption 1 holds, $\alpha = 1$, $K \ge 2$, $\kappa = \min(K-1, d-1)$, $n_A = \lfloor n/2 \rfloor$, $n_B = n - n_A$, $k \ge K-1$ and $r_{\max} \le D$, with $r_{\max}$ as defined before Theorem A′ and, for the ε-ball, $r_{\max} = \varepsilon$ and $k = c_0 n_A \varepsilon^d \ge 2K - 2$ as in Remark E.1, so that $\tau = \lceil k/2 \rceil \ge K - 1$ below and every base has $K-1$ distinct candidates; $h_\star = 2r_{\max}$, and $R_j(x)$ is the distance from x to its j-th nearest neighbour in $\mathcal{D}_A$. Under (2), or $\varepsilon = \varepsilon_0 n^{-1/(2d-\bar{\kappa})}$ for the ball, $h_\star / D \asymp (k/n)^{1/d} \asymp n^{-1/(2d-\bar{\kappa})}$ for both rules.

Truncation and the three terms. The *candidates* of a base x are the points among which its partners are drawn. For k-NN they are the k nearest neighbours of x in $\mathcal{D}_A$, and in the truncated generator one farther than $r_{\max}$ from x is replaced by an independent draw from $\mu_0|_{B(x, r_{\max})}$. For the ε-ball they are the $N(x)$ pool points in $B(x, \varepsilon)$, and in the truncated generator they are supplemented by independent draws from $\mu_0|_{B(x,\varepsilon)}$ up to $\tau := \lceil k/2 \rceil$ candidates when $N(x) < \tau$. The partners are a uniform set of $K-1$ candidates; write Φ_x° for the distribution of the synthetic point from base x, $\widehat{\nu}_X^\circ := \frac{1}{n_B} \sum_b \Phi_{X_b}^\circ$, $\bar{\nu}_A := \int \mu_0(dz)\, \Phi_z^\circ = \mathbb{E}[\widehat{\nu}_X^\circ \mid \mathcal{D}_A]$ and $\bar{\nu}^\circ := \mathbb{E}\bar{\nu}_A$. For fixed $x \in \Omega$ the number of pool points in $B(x, r_{\max})$ is binomial with mean at least $n_A c_0 r_{\max}^d$, which is $4k$ for k-NN and k for the ε-ball, so the multiplicative Chernoff bound gives $\mathbb{P}(R_k(x) > r_{\max}) \le e^{-k/2}$ and $\mathbb{P}(N(x) < \tau) \le e^{-k/8}$: a base is affected by the truncation with probability at most $p_1 := e^{-k/2}$ for k-NN and $p_1 := e^{-k/8}$ for the ball. Only affected bases change their distribution, so $d_{TV}(\widehat{\nu}_X, \widehat{\nu}_X^\circ)$ is at most the fraction of affected bases, and since $W_1 \le D\, d_{TV}$ on $\operatorname{conv}\Omega$, which has diameter D, $\mathbb{E}W_1(\widehat{\nu}_X, \widehat{\nu}_X^\circ) \le D p_1$ and, by convexity, $W_1(\bar{\nu}, \bar{\nu}^\circ) \le D p_1$, with no union bound over the bases. On the event $\mathcal{E}$ that no base is affected, of probability at least $1 - n_B p_1 \ge 1 - n p_1$, $\widehat{\nu}_X^\circ = \widehat{\nu}_X$ and every tuple has spread at most $h_\star$; $\mathcal{E}$ enters only the deviation bound. Finally

$$W_1(\widehat{\nu}_X^\circ, \mu_0) \;\le\; \underbrace{W_1(\widehat{\nu}_X^\circ, \bar{\nu}_A)}_{\text{bases given the pool}} \;+\; \underbrace{W_1(\bar{\nu}_A, \bar{\nu}^\circ)}_{\text{pool}} \;+\; \underbrace{W_1(\bar{\nu}^\circ, \mu_0)}_{\text{bias}}. \tag{56}$$

Splitting is what makes the first term a sum of independent terms: given $\mathcal{D}_A$, the bases are i.i.d. μ_0 and $\widehat{\nu}_X^\circ(Q)$ is an average of n_B i.i.d. copies of $\Phi_X^\circ(Q)$. In both rules the candidates of z are, given their number (and for k-NN given $\rho := \min\{R_{k+1}(z), r_{\max}\}$), i.i.d. $\mu_0|_{B(z,\rho)}$, with $\rho = \varepsilon$ for the ball: given a region and the number of pool points in it, those points are i.i.d. μ_0 restricted to it, and the supplementary draws have the same distribution. A uniform set of $K-1$ of them is again i.i.d.

Lemma E.2 (Small-ball bound, base independent of the pool)**.** *Let $z \sim \mu_0$ be independent of $\mathcal{D}_A$. Let the partner set σ be either a uniform set of $K-1$ candidates of z or, for k-NN, the k-th or $(k+1)$-st nearest neighbour of z together with a uniform set of $K-2$ other candidates, and let Y be the synthetic point built on $\{z\} \cup \sigma$. Then for all $t > 0$*

$$q(t) \;:=\; \mathbb{E}\Big[\sup_y \mathbb{P}\big(\|Y - y\| \le t \mid z, \sigma\big)\Big] \;\le\; C_{K,d}\Big(\frac{t}{h_\star} \wedge 1\Big)^\kappa, \tag{57}$$

with $\kappa = \min(K-1, d-1)$ and $C_{K,d}$ depending on K, d and c_f only. In particular two synthetic points of one tuple, with independent weights, satisfy $\mathbb{P}(\|Y - Y'\| \le t) \le q(t)$, the small-ball form (43) of Assumption 2, for both split generators and every $K \ge 2$.

Proof. Given the vertices, let T consist of z and $\min(K, d) - 1$ members of σ, the boundary neighbour among them if there is one, and put $B := \sum_{v \in T} U_v$. By Dirichlet aggregation $Y = BY_T + (1-B)Y_{\text{rest}}$, where Y_T is a flat Dirichlet combination of T independent of (B, Y_{rest}); $B = 1$ if $K \le d$, and $B \sim \text{Beta}(d, K-d)$ with $\mathbb{E}B^{-(d-1)} = \binom{K-1}{d-1}$ if $K > d$. With M the $d \times \kappa$ matrix of offsets of $T \setminus \{z\}$ from z, $Y_T - z = Mu$ with u uniform on $\{u \ge 0, \sum u_j \le 1\} \subset \mathbb{R}^\kappa$, of density $\kappa!$, so by Step 1 of Proposition C.4

$$\sup_y \mathbb{P}\big(\|Y - y\| \le t \mid z, \sigma\big) \;\le\; \mathbb{E}_B\Big[1 \wedge \frac{\kappa!\, \omega_\kappa\, (t/B)^\kappa}{\prod_i \sigma_i(M)}\Big].$$

Condition on z and on ρ as above; the candidates are i.i.d. $\mu_0|_{B(z,\rho)}$ (Step 2 of Proposition C.4, and Corollary C.6 for the ball; the base is now independent of the pool, which only simplifies it), except that a boundary neighbour lies on $\partial B(z,\rho)$ (for the k-th neighbour, condition on $R_k(z) \le r_{\max}$ instead). Run Gram–Schmidt on the columns of M, the boundary neighbour first: its column has norm ρ, a first column drawn from the ball has $\mathbb{P}(\|\cdot\| \le \delta\rho) \le c_f \delta^d$, and every later column is an independent draw whose conditional probability of coming within $\delta\rho$ of the span of the previous $\ell - 1$ columns is at most $C_0 \delta^{\,d-\ell+1}$, with $d - \ell + 1 \ge d - \kappa + 1 \ge 2$. Step 3 of Proposition C.4 then gives $\mathbb{E}[\rho^\kappa / \prod_i \sigma_i(M) \mid z, \rho] \le \Lambda'_{K,d}$ uniformly, and $\mathbb{E}[(h_\star/\rho)^\kappa] \le c$: for the ball $h_\star/\rho = 2$, and for k-NN Step 4, applied to R_{k+1} with n_A pool points, together with $h_\star/r_{\max} = 2$. Multiplying, and by $\mathbb{E}B^{-\kappa} \le \binom{K-1}{d-1}$, (57) holds for $t < h_\star$; for $t \ge h_\star$ it is trivial. Conditioning on Y' gives $\mathbb{P}(\|Y - Y'\| \le t) \le q(t)$. □

At $K \le d$ the aggregation is idle and $\kappa = K - 1$. Above d it discards the directions beyond $d - 1$: the exponent $d - 1$ is not sharp (the full d, with a logarithm at $K = d + 1$, is what Remark C.7 finds), but it exceeds $d - 2$, which is all the rate needs. Measured on the tuple distribution of the proof with 8×10^6 draws, the small-ball exponent of $\|Y - Y'\|$ fitted over $t \in [0.02, 0.1]\, r_{\max}$, where the smallest bin still holds 640–16500 hits, is 2.33–2.76 for $d = 3$, $K = 4, 5, 6$, and 3.12–3.74 for $d = 4$, $K = 5, 6, 8$, above $d - 1$ in every case. The local slopes at the small end, 2.46–2.87 and 3.32–3.87, sit closer to the true d, lowest at $K = d + 1$ where the logarithm is, and fall towards the large end as the probability saturates, which says nothing about the bound.

Lemma E.3 (The bases given the pool)**.** *With $\theta_{\mathrm{sm}} = h_\star/(C_Q D)$ and $\theta_\star$, $R_{d,\kappa}$ as in Definition B.2 at $S = 1/n_B$, $\mathbb{E}W_1(\widehat{\nu}^\circ_X, \bar{\nu}_A) \le C_d D \psi_d(1/n_B) \wedge c_d D\big(n_B^{-1/2}\Phi_d(\theta_\star) + R_{d,\kappa}(\theta_\star)\big)$.*

Proof. Given $\mathcal{D}_A$, $\mathrm{Var}(\widehat{\nu}^\circ_X(Q) \mid \mathcal{D}_A) \le n_B^{-1}\mathbb{E}[\Phi^\circ_X(Q)^2 \mid \mathcal{D}_A]$. Φ°_x averages the component distributions $\rho_{x,\sigma}$ over partner sets, so by Jensen $\Phi^\circ_x(Q)^2 \le \mathrm{avg}_\sigma\, \rho_{x,\sigma}(Q)^2$, and $\sum_Q \rho_{x,\sigma}(Q)^2$ is the probability that two points of one tuple share a cell, at most $\mathbb{P}(\|Y - Y'\| \le \sqrt{d}s) \le q(\sqrt{d}s)$. By Lemma E.2,

$$\mathbb{E} \sum_{Q \in \mathcal{Q}_\ell} \mathrm{Var}\left(\widehat{\nu}^\circ_X(Q) \mid \mathcal{D}_A\right) \;\le\; \frac{C_{K,d}}{n_B}\Big(\frac{\sqrt{d}D\theta_\ell}{h_\star} \wedge 1\Big)^\kappa. \tag{58}$$

Run the recursion (22) for $(\widehat{\nu}^\circ_X, \bar{\nu}_A)$ conditionally on $\mathcal{D}_A$, bound $\mathbb{E}[A_\ell \mid \mathcal{D}_A]$ by Cauchy–Schwarz over the θ_ℓ^{-d} cells, and average over $\mathcal{D}_A$ with Jensen: $\mathbb{E}A_\ell \le \theta_\ell^{-d/2}(\mathbb{E}\sum_Q \mathrm{Var})^{1/2}$. This is (33) with $S = 1/n_B$, exponent κ and no factor b_K, since distinct bases give independent components. Steps 1–2 of Section B conclude. The supplementary draws of the truncation are what keep (58) free of an additive term: a candidate copied from the base would make its simplex degenerate. For the baseline, $\Phi^\circ_X(Q) \le 1$ gives $\mathrm{Var}(\widehat{\nu}^\circ_X(Q) \mid \mathcal{D}_A) \le \bar{\nu}_A(Q)/n_B$, the variance bound of Section A.4 Step 4 with $S = 1/n_B$, and Steps 3–6 there conclude. □

Lemma E.4 (The pool)**.** *With the same θ_{sm} and with $\theta_\star$, $R_{d,\kappa}$ at $S = 1/n_A$, $\mathbb{E}W_1(\bar{\nu}_A, \bar{\nu}^\circ) \le C'_{d,K} D \psi_d(1/n_A) \wedge c'_d D\big(n_A^{-1/2}\Phi_d(\theta_\star) + R_{d,\kappa}(\theta_\star)\big)$.*

Proof. $\mathbb{E}\bar{\nu}_A(Q) = \bar{\nu}^\circ(Q)$ for every cell, so it suffices to prove (D1) in summed form with $S = 1/n_A$ and then chain as in Lemma E.3. Replace the j-th pool point X'_j by an independent copy X''_j and let Δ_j be the change of $\bar{\nu}_A$. For a base z the candidates change only if X'_j or X''_j lies within $r_{\max}$ of z. For k-NN at most two change: X'_j leaves and the $(k+1)$-st neighbour (or a supplementary draw) enters, X''_j enters and the k-th neighbour (or a supplementary draw) leaves. Partner sets avoiding the changed candidates are common to both configurations, so $\Phi^\circ_z - \Phi'^\circ_z$ is a signed sum of four measures, each $\binom{k}{K-1}^{-1}$ times the sum of $\rho_{z,\sigma}$ over the partner sets σ containing one changed candidate, the *anchor*. For the ball the number M of candidates may also change by one; removing a candidate a turns Φ°_z into Φ' with $\Phi^\circ_z - \Phi' = \frac{K-1}{M}(\mathrm{avg}_{\sigma \ni a}\rho_{z,\sigma} - \Phi')$, so besides at most four anchored measures there are at most two copies of $\frac{K-1}{M}\Phi'$, which are anchored at the base. In both rules each of these at most six measures has mass at most $\frac{2(K-1)}{k}$ per base, since $M \ge \tau \ge k/2$, and integrating in z, $\Delta_j = \sum_{r \le 6} \pm\sigma_r$ with $\sigma_r(\mathbb{R}^d) \le \frac{2(K-1)}{k}\,\mu_0\big(B(X'_j, r_{\max}) \cup$

$B(X_j'', r_{\max})\big) \le 16(K-1)c_f/n_A$, because $\mu_0(B(x, r_{\max})) \le \|f\|_\infty \omega_d r_{\max}^d \le 4c_f k/n_A$. For any measure, $\sum_Q \sigma(Q)^2 \le \sigma(\mathbb{R}^d)\, \sup_y \sigma(B(y, \sqrt{d}s))$, and

$$\sup_y \sigma_r\big(B(y,t)\big) \;\le\; \int \mu_0(dz) \binom{M}{K-1}^{-1} \sum_{\sigma \ni a} \sup_y \mathbb{P}\big(Y_{z,\sigma} \in B(y,t)\big),$$

the sum running over all partner sets for a base-anchored measure. Summed over j, every candidate of z is the anchor X_j' for at most one j, the entering point X_j'' is a candidate of an i.i.d. pool, and the other anchors change only when X_j' or X_j'' is a candidate; since $\binom{M}{K-1}^{-1} \sum_a \sum_{\sigma \ni a} = (K-1)\,\mathrm{avg}_\sigma$, each role contributes at most $2(K-1)\,q(t)$, with q taken over the partner-set distribution of that role in Lemma E.2; the anchor itself plays no role, since the supremum over y does not see where the simplex is centred. With $\sum_Q \Delta_j(Q)^2 \le 6 \sum_r \sum_Q \sigma_r(Q)^2$ and the Efron–Stein inequality,

$$\sum_{Q \in \mathcal{Q}_\ell} \mathrm{Var}\, \bar\nu_A(Q) \;\le\; \frac{1}{2} \sum_{j \le n_A} \mathbb{E} \sum_Q \Delta_j(Q)^2 \;\le\; \frac{576(K-1)^2 c_f\, C_{K,d}}{n_A} \Big(\frac{\sqrt{d} D \theta_\ell}{h_\star} \wedge 1 \Big)^\kappa. \tag{59}$$

For the baseline, $\sum_Q \Delta_j(Q)^2 \le (\sum_r \sigma_r(\mathbb{R}^d))^2 \le (96(K-1)c_f/n_A)^2$, so $\sum_Q \mathrm{Var}\, \bar\nu_A(Q) \le C/n_A$, which is the form of Section A.4 Step 4 with $S = 1/n_A$. □

The exponent in (59) is κ. It suffices for Theorem A, because the pool then enjoys exactly the (D1) of the bases, with n_A in place of n_B. The sharper two-point form, which Lemma E.9 proves for the bases, is stated for the pool as Assumption 3, and the Monte Carlo evidence for it is given after that assumption. Without the aggregation the relevant mean $\mathbb{E}[\rho^{K-1}/\prod_i \sigma_i]$, with offsets taken from a pool point, is finite and stable in the sample size for $K \le d$ (1.50, 3.5, 1.33, 2.26, 6.03, 1.63, 5.31, 16.9 for $(d,K) = (3,2),(3,3),(4,2),(4,3),(4,4),(6,3),(6,5),(6,6)$ over 10^4 to 10^6 draws) and heavy-tailed at $K = d+1$, with small-ball exponent 0.99–1.01; this is why $K > d$ goes through a d-point sub-tuple.

For the ε-ball the bias needs nothing new: the candidates of z are i.i.d. $\mu_0|_{B(z,\varepsilon)}$ whatever their number, so $\bar\nu^\circ$ is exactly the population distribution of Proposition 4.1 with radius ε, and $W_1(\bar\nu^\circ, \mu_0) \le (\frac{K-1}{(d+2)(K+1)} \|\nabla f\|_{L^1} + C_\partial)\varepsilon^2 + o(\varepsilon^2)$. For k-NN the radius is random and adapts to the density.

Lemma E.5 (k-NN mixing bias)**.** *With C_∂' the collar contribution of Proposition 4.1,*

$$\begin{aligned} W_1(\bar\nu^\circ, \mu_0) \;\le\; & \Big(\frac{K-1}{(d+2)(K+1)} \Big(1 + \frac{2}{dK}\Big) \frac{\|f^{-2/d} \nabla f\|_{L^1}}{\omega_d^{2/d}} + C_\partial' \Big) \Big(\frac{k+1}{n_A} \Big)^{2/d} \\ & + o\Big(\Big(\frac{k}{n} \Big)^{2/d} \Big) + D e^{-k/2}. \end{aligned} \tag{60}$$

Proof. The untruncated population distribution $\bar\nu := \int \mu_0(dz) \mathbb{E}\Phi_z$ differs from $\bar\nu^\circ$ in total variation by at most $\sup_z \mathbb{P}(R_k(z) > r_{\max}) \le e^{-k/2}$, which costs $De^{-k/2}$. Put $F_z(r) := \mu_0(B(z,r))$. Given z, the variable $q := F_z(R_{k+1}(z))$ has the $\mathrm{Beta}(k+1, n_A - k)$ distribution, which does not depend on z, and given q the k nearest neighbours are i.i.d. $\mu_0|_{B(z,\rho_q(z))}$ with $\rho_q(z) := F_z^{-1}(q)$, the radius of the ball of mass q. So $\bar\nu = \mathbb{E}_q \bar\nu^{(q)}$, where $\bar\nu^{(q)}$ draws partners from the mass-q ball, and $W_1(\bar\nu, \mu_0) \le \mathbb{E}_q W_1(\bar\nu^{(q)}, \mu_0)$ by convexity. For fixed q the field ρ_q is C^2 away from $\partial\Omega$ (implicit functions, $\partial_r F_z > 0$, $f \in C^2$), and the proof of Proposition 4.1 applies with $h_\star$ replaced by $\rho_q(z)$: the moments (51) hold pointwise with $h_\star^2 \to \rho_q(z)^2$, the expansion being local in z, and in $V_0 = fb - \frac{1}{2}\nabla \cdot (fM)$ the divergence now also hits ρ_q^2,

$$V_0 \;=\; \frac{K-1}{K(K+1)(d+2)} \big(K \rho_q^2 \nabla f - f \nabla \rho_q^2 \big) + o(\rho_q^2).$$

The first-order term of f integrates to zero over a ball, so $F_z(r) = \omega_d r^d (f(z) + O(r^2))$, $\rho_q^d = \frac{q}{\omega_d f}(1 + O(\rho_q^2))$ and $\nabla \rho_q^2 = -\frac{2}{d}\rho_q^2 \nabla \log f + O(\rho_q^3)$, whence $V_0 = \frac{K-1}{(d+2)(K+1)}(1 + \frac{2}{dK})\rho_q^2 \nabla f + o(\rho_q^2)$. The collar of width ρ_q along $\partial\Omega$ adds $C_\partial' q^{2/d}$ as in Proposition 4.1. Finally, for $d \ge 2$, $\mathbb{E}q^{2/d} \le (\mathbb{E}q)^{2/d} \le ((k+1)/n_A)^{2/d}$ by Jensen, as $2/d \le 1$; at $d = 1$, $\mathbb{E}q^2 = \frac{(k+1)(k+2)}{(n_A+1)(n_A+2)} \le ((k+2)/n_A)^2$, and (60) holds with $k+2$ in place of $k+1$. □

The factor $1+2/(dK)$ is the price of a radius that adapts to the density: the k-NN ball is wider where f is small, and the resulting drift points down the density gradient, adding to the fixed-radius term. On the circle ($d=1$) with $f=1+0.6\cos 2\pi x$ the measured bias of the mass-q generator is 0.94, 0.98, 0.98 of (60) at $K=2$ for $q=0.10$, 0.07, 0.05, and 0.94–0.99 at $K=3,5$; with the random q of the k-NN rule ($k=20,40$) it is within 1%. The fixed-radius formula at the local radius, which drops $\nabla\rho_q^2$, is off by 1.77–1.92 at $K=2$, approaching $1+2/K=2$.

Lemma E.6 (Bounded differences). *Let $F:=W_1(\widehat{\nu}_X^\circ,\mu_0)$. For every $\delta\in(0,1)$, with probability at least $1-\delta$, $|F-\mathbb{E}F|\le 20(K-1)c_f D\sqrt{\log(2/\delta)/n}$.*

Proof. Reveal $\mathcal{D}_A$ point by point, then $\mathcal{D}_B$, and apply the Azuma–Hoeffding inequality to the Doob martingale of F; it needs, for each coordinate, a bound c_j on the range of the conditional expectation as that coordinate varies. Changing a base changes $\widehat{\nu}_X^\circ$ by mass $1/n_B$, so $c_j=D/n_B$ there, since $W_1\le D\,d_{TV}$ on $\operatorname{conv}\Omega$. Changing a pool point changes Φ_z° only for $z\in B(X_j',r_{\max})\cup B(X_j'',r_{\max})$, and there by at most $2(K-1)/M$ in total variation, M the number of candidates: by the decomposition in the proof of Lemma E.4, at most two candidates change and each is met by a uniform partner set with probability $(K-1)/M$. For k-NN, $M=k$ and the affected bases have μ_0-mass at most $8c_fk/n_A$; for the ball, $M\ge k/2$ and the mass is at most $2c_fk/n_A$. The conditional expectation integrates the bases, so for every configuration of the pool the change is at most $16(K-1)c_fD/n_A$ in both rules. Azuma–Hoeffding gives $\mathbb{P}(|F-\mathbb{E}F|\ge t)\le 2\exp(-2t^2/\sum_j c_j^2)$ with $\sum_j c_j^2=D^2(256(K-1)^2c_f^2/n_A+1/n_B)\le 770(K-1)^2c_f^2D^2/n$, as $n_A\ge n/3$, $n_B\ge n/2$ and $(K-1)c_f\ge 1$; and $\sqrt{770/2}<20$. □

Proof of Theorem A′ for (G2), and of Remark E.1. (i) By the triangle inequality, $W_1(\widehat{\nu}_X,\bar\nu)\le W_1(\widehat{\nu}_X,\widehat{\nu}_X^\circ)+W_1(\widehat{\nu}_X^\circ,\bar\nu_A)+W_1(\bar\nu_A,\bar\nu^\circ)+W_1(\bar\nu^\circ,\bar\nu)$. The first and last terms are at most Dp_1 in expectation, the additional term of (i), and Lemmas E.3 and E.4 bound the middle ones with the exponent κ of Assumption 2 at every arity; as $n_B\ge n_A$ and both bounds increase with S, both forms of (i) hold with $S=1/n_A$. (ii) On $\mathcal{E}$, $W_1(\widehat{\nu}_X,\mu_0)=F$, and $|\mathbb{E}W_1(\widehat{\nu}_X,\mu_0)-\mathbb{E}F|\le \mathbb{E}W_1(\widehat{\nu}_X,\widehat{\nu}_X^\circ)\le Dp_1$. If $p_n=np_1\ge 1$ there is nothing to prove; otherwise $Dp_1<D/n\le D\sqrt{\log 2/n}$ for $n\ge 2$, and Lemma E.6 on $\mathcal{E}$, whose complement has probability at most p_n, gives (ii), the additive term turning the constant 20 into 21. (iii) $\beta_1\le W_1(\bar\nu^\circ,\mu_0)+Dp_1$, which is $O(\varepsilon^2)$ for the ball by Proposition 4.1 and $O((k/n)^{2/d})$ for k-NN by Lemma E.5; both are $O(h_\star^2)$. Under (2), or $\varepsilon=\varepsilon_0 n^{-1/(2d-\bar\kappa)}$ for the ball, both rules have $k\asymp n^{(d-\bar\kappa)/(2d-\bar\kappa)}$, $r_{\max}\le D$ and $p_1\le n^{-2/(2d-\bar\kappa)}$ for $n\ge n_0$, $\theta_{\mathrm{sm}}\asymp n^{-1/(2d-\bar\kappa)}$ and $\theta_{\mathrm{samp}}\asymp n^{-1/d}$, and the ratio $G=\theta_{\mathrm{sm}}/\theta_{\mathrm{samp}}\asymp k^{1/d}/C_Q$ exceeds one, so Proposition B.4 (Step 4 of Section B) applies to (54). For $K\ge d$, $\kappa=d-1>d-2$ and the fluctuation is $\lesssim n^{-1/2}\theta_{\mathrm{sm}}^{1-d/2}\asymp n^{-2/(d+2)}$; for $K=d-1$ it carries the extra factor $1+\frac{d}{2}\log G\lesssim\log n$; for $K\le d-2$ it is $\lesssim\theta_{\mathrm{samp}}^{d/(d-\kappa)}\theta_{\mathrm{sm}}^{-\kappa/(d-\kappa)}\asymp n^{-1/(d-\kappa)+\kappa/((d-\kappa)(2d-\kappa))}=n^{-2/(2d-\kappa)}$. The bias is $O(h_\star^2)=O(n^{-2/(2d-\bar\kappa)})$ and the remainders are of lower order, which with $\mathbb{E}W_1(\widehat{\nu}_X,\mu_0)\le\mathbb{E}W_1(\widehat{\nu}_X,\bar\nu)+\beta_1$ is (iii). □

Proofs of the corollaries. *The ratio* (4). Any coupling moves the μ_0-mass at y to a sample point, so for every sample $W_1(\hat\mu_n,\mu_0)\ge\int\operatorname{dist}(y,\{X_i\})\,\mu_0(dy)\ge r\big(1-n\|f\|_\infty\omega_d r^d\big)$, and $r=(2n\|f\|_\infty\omega_d)^{-1/d}$ gives $W_1(\hat\mu_n,\mu_0)\ge\frac12(2\|f\|_\infty\omega_d)^{-1/d}n^{-1/d}$. Divide Theorem A(i) by it and use $\frac{2}{2d-\bar\kappa}-\frac1d=\frac{\bar\kappa}{d(2d-\bar\kappa)}$. The exponent increases with $\bar\kappa$, hence with K up to $K=d-1$; $K=1$ is Proposition G.3. *Corollary A.13.* The points of a tuple are i.i.d. μ_0, so Steps 1 and 3 of Proposition C.4 apply with the ball $B(x,\rho)$ replaced by $\operatorname{supp}\mu_0$: the δD-neighbourhood of an $(\ell-1)$-dimensional affine subspace meets it in volume at most $c_dD^d\delta^{\,d-\ell+1}$, so the conditional probability is at most $c_df_{\max}D^d\delta^{\,d-\ell+1}$, and Assumption 2 holds with $h_\star=D$ and $C_{K,d}$ depending on d, K and $f_{\max}D^d$; Step 4 is not needed. For $K>d$ the aggregation of Lemma E.2 onto a d-point sub-tuple gives it with $\kappa=d-1$. Then $\theta_{\mathrm{sm}}=1/C_Q$ is fixed, and $S_\pi=1/n$ gives $\mathbb{E}S\le 2/n$ for $n_\pi\ge n$ (Section G Step 0), so $\theta_{\mathrm{samp}}<\theta_{\mathrm{sm}}$ for $n\ge n_0$. Theorem A′(i) and Proposition B.4 give, conditionally on the draw, $\sqrt{S}\,\Phi_d(\theta_{\mathrm{sm}})\asymp\sqrt{S}$ for $K\ge d$, with the extra factor $\log(1/S)$ at $K=d-1$, and $\theta_{\mathrm{samp}}^{d/(d-\kappa)}\theta_{\mathrm{sm}}^{-\kappa/(d-\kappa)}\asymp S^{1/(d-\kappa)}$ for $K\le d-2$; each is a concave function of S in this range, so averaging over the draw replaces S by $\mathbb{E}S\le 2/n$. *Corollary 3.2.* Proposition A.11 and Theorem A. □

Corollary 3.1. Theorem A′ bounds the synthetic distribution $\widehat{\nu}_X$, in which the Dirichlet weights, and in (G2) the partner sets, are integrated out. *Lower bounds.* Let γ be an optimal plan between μ_0 and M_λ. The atoms of $\hat{\mu}_n$ receive μ_0-mass at least $1-\lambda$, and the μ_0-mass within distance r of them is at most $n\|f\|_\infty \omega_d r^d$; at $r = ((1-\lambda)/(2n\|f\|_\infty\omega_d))^{1/d}$ at least $(1-\lambda)/2$ of it travels at least r, which gives $W_1(M_\lambda,\mu_0) \geq c_{\mathrm{q}}(1-\lambda)^{1+1/d}n^{-1/d}$. The same argument with all $n+N$ atoms of M_λ and mass one is the proof of (4) and gives $c_{\mathrm{q}}(n+N)^{-1/d}$. *Upper bound.* By joint convexity, $W_1(M_\lambda,\mu_0) \leq (1-\lambda)W_1(\hat{\mu}_n,\mu_0) + \lambda W_1(\widehat{\nu}_{X,N},\mu_0)$; given the data the N points are i.i.d. $\widehat{\nu}_X$ on a set of diameter D, so Steps 2–6 of Section A.4 with $S = 1/N$ give $\mathbb{E}[W_1(\widehat{\nu}_{X,N},\widehat{\nu}_X) \mid X] \leq C_d D \psi_d(1/N)$, and the same with $S = 1/n$ bounds $\mathbb{E}W_1(\hat{\mu}_n,\mu_0)$. *The two regimes.* At the tuned parameters the upper bound is of order $n^{-2/(2d-\bar{\kappa})}\ell_n$ once $N^{-1/d}$ and $(1-\lambda)n^{-1/d}$ fall below it, which is the condition stated; $N \gtrsim n^{2d/(2d-\bar{\kappa})}$ is $n^{2d/(d+2)}$ from $K = d-1$ on ($n^{1.2}$ at $d=3$, $n^{5/3}$ at $d=10$). Conversely the lower bound is of order $n^{-1/d}$ when $N \asymp n$ or $1-\lambda$ is bounded below. In particular giving each of N synthetic points the weight n/N next to the n real points, $\lambda = \frac{1}{2}$, does not beat $n^{-1/d}$ however large N is. SMOTE to class parity at imbalance ratio ρ has $N = (\rho-1)n$ and $1-\lambda = 1/\rho$, with n the minority count, so it keeps the rate iff $\rho \gtrsim n^{\bar{\kappa}/(2d-\bar{\kappa})}$; at a bounded ratio both conditions fail, and only a constant-factor gain remains possible. The same holds in (G1), where a synthetic sample draws one weight vector per tuple, and in Corollary 3.2, where training on M_λ replaces $W_1(\widehat{\nu}_X,\mu_0)$ by $W_1(M_\lambda,\mu_0)$. mixup, which draws fresh points at every step, trains on $\widehat{\nu}_X$ in the limit. The rate experiments of Section I use synthetic clouds alone, above the threshold. □

Lemma E.7 (A lower bound on the k-NN bias)**.** *In (G2) on $\mathbb{T}^d$, with $\mathrm{Dir}(\alpha,\dots,\alpha)$ weights for any $\alpha > 0$, let $f \in C^2$ with $\nabla f \not\equiv 0$. There are $r_1 > 0$ and $c > 0$, depending only on d, K and f, such that $r_{\max} \leq r_1$ implies $W_1(\bar{\nu}^\circ,\mu_0) \geq c\, r_{\max}^2$.*

Proof. As in the proof of Lemma E.8, given $q \sim \mathrm{Beta}(k+1, n_A-k)$ the partners of z are i.i.d. $\mu_0|_{B(z,r(z))}$ with $r := \min\{\rho_q, r_{\max}\}$, a Lipschitz radius field bounded by $r_{\max}$; let ν_q be the resulting distribution, so that $\bar{\nu}^\circ = \mathbb{E}_q\nu_q$. The function $g := f/\|\nabla f\|_\infty$ is 1-Lipschitz, so $W_1(\bar{\nu}^\circ,\mu_0) \geq \mathbb{E}_q \int g\, d(\nu_q - \mu_0)$. The moments (51) hold pointwise with $h_\star$ replaced by $r(z)$ and $\mathbb{E}U_j^2 = \lambda/K$, $\lambda := (\alpha+1)/(K\alpha+1) \leq 1$, with remainders $O(r_{\max}^3)$. A second-order Taylor expansion of g, whose remainder is at most $\omega_{D^2 g}(r_{\max})r_{\max}^2$, and one integration by parts on $\mathbb{T}^d$, legitimate as r^2 is Lipschitz, give, uniformly in q,

$$\int g\, d(\nu_q - \mu_0) \;=\; \frac{K-1}{K(d+2)\|\nabla f\|_\infty} \int_{\mathbb{T}^d} \Big((1-\tfrac{\lambda}{2}) r^2 |\nabla f|^2 - \tfrac{\lambda}{2}\, f\, \nabla f \cdot \nabla r^2 \Big) \;+\; o(r_{\max}^2).$$

Where $\rho_q < r_{\max}$, $\nabla r^2 = \nabla \rho_q^2 = -\frac{2}{d}\rho_q^2 \nabla \log f + O(\rho_q^3)$ (proof of Lemma E.5), so $f\,\nabla f \cdot \nabla r^2 \leq O(r_{\max}^3)$; elsewhere $r = r_{\max}$ and $\nabla r^2 = 0$. The integrand is therefore at least $\frac{1}{2}r^2|\nabla f|^2 - O(r_{\max}^3)$. Since $\mu_0(B(z,r)) \leq \omega_d f_{\max} r^d$, the radius satisfies $\rho_q^d \geq q/(\omega_d f_{\max})$, so on $\{q \geq k/(2n_A)\}$ we have $r \geq c_1 r_{\max}$ everywhere, with $c_1^d := c_0/(8\omega_d f_{\max}) \leq 1$; and $\mathbb{P}(q \geq k/(2n_A)) = \mathbb{P}(\mathrm{Bin}(n_A, k/(2n_A)) \leq k) \geq \frac{1}{2}$ by Markov's inequality. Averaging over q,

$$W_1(\bar{\nu}^\circ,\mu_0) \;\geq\; \frac{(K-1)\, c_1^2\, \|\nabla f\|_{L^2}^2}{4K(d+2)\|\nabla f\|_\infty}\, r_{\max}^2 \;-\; o(r_{\max}^2),$$

and a small r_1 makes the remainder at most half the first term. □

The truncated distribution Φ_x° is carried by $B(x, r_{\max})$, so Theorem B applies to the truncated generator with $h = r_{\max}$, and $\mathbb{E}W_1(\widehat{\nu}_X,\mu_0) \geq \mathbb{E}W_1(\widehat{\nu}_X^\circ,\mu_0) - De^{-k/2}$ transfers it to k-NN SMOTE once $k \gtrsim \log n$.

Proof of Theorem B.

Proof. Bias. By joint convexity of W_1, $\mathbb{E}W_1(\widehat{\nu}_X,\mu_0) \geq W_1(\bar{\nu},\mu_0) = \beta_1$. *Fluctuation.* With $F := \mathbb{E}W_1(\widehat{\nu}_X,\bar{\nu})$ the triangle inequality gives $\mathbb{E}W_1(\widehat{\nu}_X,\mu_0) \geq F - \beta_1$, so $\mathbb{E}W_1(\widehat{\nu}_X,\mu_0) \geq \frac{1}{2}\max\{\beta_1, F\}$. Put $r := C\max\{h, n^{-1/d}\}$, with C fixed below and h_0, n_0 such that $r \leq \frac{1}{4}$; let c_Q run over the centres of a grid of side r and $\varphi_Q(y) := (r/2 - |y - c_Q|)_+$. These tents are 1-Lipschitz with disjoint supports, so $\sum_Q \varepsilon_Q \varphi_Q$ is 1-Lipschitz for every choice of signs, and choosing them

by the data, $W_1(\widehat{\nu}_X, \bar{\nu}) \geq \sum_Q |a_Q + b_Q|$ with $a_Q := \int \varphi_Q \, d(\widehat{\nu}_X - \bar{\nu}_A)$, $b_Q := \int \varphi_Q \, d(\bar{\nu}_A - \bar{\nu})$ and $\bar{\nu}_A := \mathbb{E}[\widehat{\nu}_X \mid \mathcal{D}_A]$. Given $\mathcal{D}_A$, b_Q is fixed and $a_Q = n_B^{-1} \sum_b (\psi_Q(X_b) - \mathbb{E}\psi_Q(X_b))$ with $\psi_Q(x) := \int \varphi_Q \, d\Phi_x$, and for a mean-zero a, $\mathbb{E}|a+b| \geq \max\{|b|, \mathbb{E}|a| - |b|\} \geq \frac{1}{2}\mathbb{E}|a|$. As Φ_x is carried by $B(x,h)$, $|\psi_Q - \varphi_Q| \leq h$, and $\psi_Q - \varphi_Q$ vanishes outside $B(c_Q, r/2 + h)$; with $p_Q := \mu_0(B(c_Q, r/2)) \asymp r^d$, $\operatorname{Var} \varphi_Q(X) \geq c_1 r^2 p_Q$ since $f_{\min} \leq f \leq f_{\max}$, so the standard deviation of $\psi_Q(X)$ is at least $(\sqrt{c_1}\, r - c_2 h)\sqrt{p_Q} \geq \frac{1}{2}\sqrt{c_1}\, r \sqrt{p_Q}$ once $C \geq 2c_2/\sqrt{c_1}$. The summands are bounded by r and $n_B p_Q \geq c_3 C^d \geq 1$, so the fourth moment of a_Q is at most a constant times the square of its variance, and Hölder's inequality, $\mathbb{E}|a| \geq (\mathbb{E}a^2)^{3/2}(\mathbb{E}a^4)^{-1/2}$, gives $\mathbb{E}|a_Q| \geq c_4 n^{-1/2} r \sqrt{p_Q}$. Summing over the $\asymp r^{-d}$ cells, $F \geq c_5 n^{-1/2} r^{1-d/2}$. *Balance.* If $h \geq n^{-1/d}$, $F \geq c_5 C^{1-d/2} n^{-1/2} h^{1-d/2}$, and if $h < n^{-1/d}$, $F \geq c_5 C^{1-d/2} n^{-1/d}$. As $h^{1-d/2} \leq n^{-(1-d/2)/d}$ for $h \geq n^{-1/d}$, both cases give $F \geq c_5 C^{1-d/2} \min\{n^{-1/2} h^{1-d/2}, n^{-1/d}\}$, which with $\beta_1 \geq c_\beta h^2$ is the first inequality, $c = \frac{1}{2}\min\{c_\beta, c_5 C^{1-d/2}\}$. For the second, $h^2 \geq n^{-2/(d+2)}$ if $h \geq n^{-1/(d+2)}$; otherwise $n^{-1/2} h^{1-d/2} \geq n^{-2/(d+2)}$, and $n^{-1/d} \geq n^{-2/(d+2)}$; the two terms cross at $h \asymp n^{-1/(d+2)}$. □

The fluctuation part uses only that synthetic points stay within h of their base, which is why it holds at every arity and weight distribution and does not see κ; what this means for optimality is discussed after Theorem B.

Theorem F.1 for (G2). The rest of this appendix proves Theorem F.1 for (G2), with the truncated distributions and the event $\mathcal{E}$ above. On $\mathbb{T}^d$ the population distribution of the truncated generator is a local displacement of μ_0 with a flux of order $h_\star^2$.

Lemma E.8 (The truncated generator on the torus)**.** *In (G2) on $\mathbb{T}^d$ there is $r_1 \in (0, \frac{1}{8}]$, depending on d, K and f, such that if $r_{\max} \leq r_1$, then $\bar{\nu}^\circ$ has a density g with $|g - f| \leq C r_{\max} \leq f_{\min}/2$, and $\bar{\nu}^\circ - \mu_0 = -\nabla \cdot J$ for a vector field J with $|J| \leq C r_{\max}^2$ everywhere, C depending on d, K and f.*

Proof. As in the proof of Lemma E.5, let $q \sim \mathrm{Beta}(k+1, n_A - k)$ and let $\rho_q(z)$ be the radius of the ball of mass q about z. Given q, the candidates of z are i.i.d. $\mu_0|_{B(z,r(z))}$ with $r := \min\{\rho_q, r_{\max}\}$: when $\rho_q(z) > r_{\max}$, the candidates inside $B(z, r_{\max})$ are i.i.d. μ_0 restricted to it given their number, and the truncation redraws the others from the same distribution. So $\bar{\nu}^\circ = \mathbb{E}_q \nu_q$, where ν_q is the distribution of $Y = z + r(z)V$ with $z \sim \mu_0$, $V = \sum_{j \geq 2} U_j V_j$, and V_j i.i.d. with density $p_z(v) = r^d f(z + rv)/\mu_0(B(z,r))$ on the unit ball, $r = r(z)$. Differentiating $\mu_0(B(z, \rho_q(z))) = q$ gives $|\nabla \rho_q| \leq C\rho_q$, so r is Lipschitz with constant $C r_{\max}$ and $r(z')/r(z) = 1 + O(|z - z'|)$; and $|\nabla_z \log p_z(v)| \leq C$, the normaliser being the constant q where $r = \rho_q$. The density $\tilde{p}(\cdot \mid z)$ of V then has $|\nabla_z \log \tilde{p}(v \mid z)| \leq C(K-1)$. For $s \in [0,1]$ and fixed v, the map $z \mapsto z + sr(z)v$ is a bijection of $\mathbb{T}^d$ with Jacobian determinant $1 + sv \cdot \nabla r = 1 + O(r_{\max})$ once $C r_{\max} \leq \frac{1}{2}$, and its inverse $z_{s,v}(x)$ lies within $r(z_{s,v}(x))$ of x. For smooth ϕ, $\int \phi \, d(\nu_q - \mu_0) = \mathbb{E} \int_0^1 \nabla\phi(z + sr(z)V) \cdot r(z) V \, ds$, so $\nu_q - \mu_0 = -\nabla \cdot J_q$ with

$$J_q(x) \;=\; \int_0^1 \!\! \int \frac{f(z)\, \tilde{p}(v \mid z)\, r(z)\, v}{1 + s\, v \cdot \nabla r(z)} \, dv \, ds, \qquad z = z_{s,v}(x).$$

Replacing z by x in f, $\tilde{p}$, r and the determinant changes the integrand by a relative $O(r_{\max})$, hence $J_q(x)$ by $O(r_{\max}^2)$, and what remains is $f(x) r(x) \, \mathbb{E}[V \mid z = x] = O(r_{\max}^2)$, because $\mathbb{E}[V_j \mid z = x] = r \nabla f(x)/((d+2) f(x)) + O(r^2)$. The same computation at $s = 1$, without the factor $r(z)v$, gives the density of ν_q as $f(x)(1 + O(r_{\max}))$. Average over q, with $J := \mathbb{E}_q J_q$, and take r_1 small. □

Proof of Theorem F.1 for (G2). On a set of diameter D, $W_2^2 \leq D^2 d_{TV}$, so by Jensen and the truncation paragraph $\mathbb{E} W_2(\widehat{\nu}_X, \widehat{\nu}_X^\circ) \leq D(\mathbb{E} d_{TV}(\widehat{\nu}_X, \widehat{\nu}_X^\circ))^{1/2} \leq D p_1^{1/2}$ and $W_2(\bar{\nu}, \bar{\nu}^\circ) \leq D p_1^{1/2}$, which is $D e^{-k/4}$ for k-NN and $D e^{-k/16}$ for the ball.

(i), first bound. Step 2 of Section F holds for any pair of measures, and Step 3 uses only the summed variance of the cell masses. The baselines of Lemmas E.3 and E.4, $\sum_Q \operatorname{Var}(\widehat{\nu}_X^\circ(Q) \mid \mathcal{D}_A) \leq 1/n_B$ and $\sum_Q \operatorname{Var} \bar{\nu}_A(Q) \leq C/n_A$, and Steps 3–6, conditionally on $\mathcal{D}_A$ for the first, bound $\mathbb{E} W_2(\widehat{\nu}_X^\circ, \bar{\nu}_A)$

and $\mathbb{E}W_2(\bar\nu_A,\bar\nu^\circ)$ by a constant times $D\varphi_d(1/n_A)$; the triangle inequality and the remainders above conclude.

(i), refined bound. Now $K\ge d$, so $\kappa=d-1>d-2$ in Lemma E.2, whose proof is local and holds on $\mathbb{T}^d$ verbatim: a tuple lies in $B(z,r_{\max})$ with $r_{\max}\le\frac18$, where the chart is an isometry and periodic and chart distances between synthetic points agree. Let G be the mean-zero Green's function of $\mathbb{T}^d$, so that $\|\sigma\|^2_{\dot H^{-1}}=\iint G(x-y)\,\sigma(dx)\,\sigma(dy)$ and $G(x)\le c_d|x|^{2-d}+c_d'$ (Proposition C.5). *Bases.* Given $\mathcal{D}_A$, $\widehat\nu_X^\circ-\bar\nu_A$ is an average of n_B i.i.d. centred measures, so, by Jensen over partner sets as in Lemma E.3 and the identity $|\hat\rho(\xi)|^2=\mathbb{E}e^{-2\pi i\xi\cdot(Y-Y')}$ of Proposition C.5, $\mathbb{E}\|\widehat\nu_X^\circ-\bar\nu_A\|^2_{\dot H^{-1}}\le n_B^{-1}\mathbb{E}[G(Y-Y')]$, with Y,Y' two synthetic points of one tuple on an independent base. By Lemma E.2 and $\mathbb{E}W^{2-d}=(d-2)\int_0^\infty t^{1-d}\mathbb{P}(W\le t)\,dt$,

$$\mathbb{E}\|Y-Y'\|^{2-d}\;\le\;(d-2)\int_0^\infty t^{1-d}q(t)\,dt\;\le\;C_{K,d}\Big(\frac{d-2}{\kappa-d+2}+1\Big)h_\star^{2-d}. \tag{61}$$

Pool. The Efron–Stein inequality, applied to the real and imaginary parts of each Fourier coefficient, gives $\mathbb{E}\|\bar\nu_A-\bar\nu^\circ\|^2_{\dot H^{-1}}\le\frac12\sum_j\mathbb{E}\|\Delta_j\|^2_{\dot H^{-1}}$, with $\Delta_j=\sum_{r\le6}\pm\sigma_r$ as in the proof of Lemma E.4. For a measure $\sigma\ge0$, $\|\sigma\|^2_{\dot H^{-1}}\le\sigma(\mathbb{T}^d)\sup_y\int G(y-x)\,\sigma(dx)$ and $\int\|y-x\|^{2-d}\sigma(dx)=(d-2)\int_0^\infty t^{1-d}\sigma(B(y,t))\,dt$, so $\sigma_r(\mathbb{T}^d)\le16(K-1)c_f/n_A$ and the bound $\sum_j\mathbb{E}\sup_y\sigma_r(B(y,t))\le2(K-1)q(t)$ of that proof give, with (61), $\mathbb{E}\|\bar\nu_A-\bar\nu^\circ\|^2_{\dot H^{-1}}\le c\,(K-1)^2c_fC_{K,d}\,n_A^{-1}h_\star^{2-d}$, the constant c_d' contributing only $O(n_A^{-1})$. Hence

$$\mathbb{E}\|\widehat\nu_X^\circ-\bar\nu^\circ\|^2_{\dot H^{-1}}\;\le\;c_{d,K}\,n_A^{-1}h_\star^{2-d}. \tag{62}$$

By Lemma E.8, $\bar\nu^\circ$ has density at least $f_{\min}/2$, so Peyre's inequality with reference $\bar\nu^\circ$, Jensen and the remainders above give (70) with $\bar f=f_{\min}$, plus the additional term.

(ii) For $F:=W_2^2(\widehat\nu_X^\circ,\mu_0)$ the proof of Lemma E.6 applies with D^2 in place of D, since $|W_2^2(\mu,\mu_0)-W_2^2(\mu',\mu_0)|\le D^2d_{TV}(\mu,\mu')$ (keep the optimal plan on the common part of μ and μ' and pair the rest arbitrarily): with probability at least $1-\delta$, $|F-\mathbb{E}F|\le20(K-1)c_fD^2\sqrt{\log(2/\delta)/n}$, and integrating the tail, $\mathbb{E}|F-\mathbb{E}F|\le35(K-1)c_fD^2n^{-1/2}$. Put $W:=\sqrt F$. Since $|\sqrt a-\sqrt b|\le\sqrt{|a-b|}$, $|W-(\mathbb{E}F)^{1/2}|\le|F-\mathbb{E}F|^{1/2}$, and $0\le(\mathbb{E}F)^{1/2}-\mathbb{E}W\le(\operatorname{Var}W)^{1/2}\le(\mathbb{E}|F-\mathbb{E}F|)^{1/2}$. With $\log(2/\delta)\ge\log2$ this gives $|W-\mathbb{E}W|\le(\sqrt{20}+\sqrt{35}(\log2)^{-1/4})\sqrt{(K-1)c_f}\,D(\log(2/\delta)/n)^{1/4}$, and the bracket is 10.96. On $\mathcal{E}$, $W_2(\widehat\nu_X,\mu_0)=W$, and $|\mathbb{E}W_2(\widehat\nu_X,\mu_0)-\mathbb{E}W|\le\mathbb{E}W_2(\widehat\nu_X,\widehat\nu_X^\circ)\le Dp_1^{1/2}$. If $p_n=np_1\ge1$ there is nothing to prove; otherwise $Dp_1^{1/2}<Dn^{-1/2}\le0.93\,D(\log(2/\delta)/n)^{1/4}$ for $n\ge2$, which gives (ii) with the constant $12>10.96+0.93$.

(iii) Peyre's inequality with reference μ_0, of density at least $f_{\min}$, and $\|\nabla\cdot J\|_{\dot H^{-1}}\le\|J\|_{L^2}$ give, by Lemma E.8, $W_2(\bar\nu^\circ,\mu_0)\le2f_{\min}^{-1/2}\|J\|_{L^2}\le Cr_{\max}^2$, so $\beta_2\le Cr_{\max}^2+Dp_1^{1/2}=O(h_\star^2)$. Likewise $W_2(\widehat\nu_X^\circ,\mu_0)\le2f_{\min}^{-1/2}(\|\widehat\nu_X^\circ-\bar\nu^\circ\|_{\dot H^{-1}}+\|J\|_{L^2})$, and $(\mathbb{E}W_2^2(\widehat\nu_X,\widehat\nu_X^\circ))^{1/2}\le Dp_1^{1/2}$, so by (62)

$$\big(\mathbb{E}W_2^2(\widehat\nu_X,\mu_0)\big)^{1/2}\;\le\;2f_{\min}^{-1/2}\big(c_{d,K}^{1/2}\,n_A^{-1/2}h_\star^{1-d/2}+Cr_{\max}^2\big)+Dp_1^{1/2}.$$

At $k=\lceil n^{2/(d+2)}\rceil$, $h_\star=2r_{\max}\asymp n^{-1/(d+2)}$ and both terms are of order $n^{-2/(d+2)}$; n_0 makes $Dp_1^{1/2}\le n^{-2/(d+2)}$ and $r_{\max}\le r_1$. The ε-ball rule of Remark E.1 is covered by the same proof, with $r\equiv\varepsilon$ in Lemma E.8. □

Proof of Corollary F.5. Bias and truncation are as in (iii): $\beta_2\le Cr_{\max}^2+Dp_1^{1/2}$ by Lemma E.8, which holds at every K, and $(\mathbb{E}W_2^2(\widehat\nu_X,\widehat\nu_X^\circ))^{1/2}\le Dp_1^{1/2}$. For the fluctuation, run (73) for $(\widehat\nu_X^\circ,\bar\nu_A)$, conditionally on $\mathcal{D}_A$, and for $(\bar\nu_A,\bar\nu^\circ)$, down to a free finest level θ_L, with $\mathbb{E}A_\ell$ from (58) and (59) as in the proofs of Lemmas E.3 and E.4. As $2^{-2(\ell-1)}=4\theta_\ell^2$ and $n_A\asymp n$, each pair has

$$\mathbb{E}W_2^2\;\le\;dD^2\theta_L^2+cD^2n^{-1/2}\sum_{\theta_L\le\theta_\ell\le1}\Big(\frac{\theta_\ell}{\theta_{\rm sm}}\wedge1\Big)^{\kappa/2}\theta_\ell^{2-d/2}.$$

For $d\ge5$ the levels above $\theta_{\rm sm}$ sum to order $n^{-1/2}\theta_{\rm sm}^{2-d/2}$. Those below sum to $\Lambda\sum\theta_\ell^{2-(d-\kappa)/2}$, $\Lambda:=n^{-1/2}\theta_{\rm sm}^{-\kappa/2}$, which is of the order of its finest term if $d-\kappa>4$, of its coarsest if $d-\kappa<4$,

and $\log(\theta_{\rm sm}/\theta_L)$ times either at $d-\kappa=4$; here $\theta_{\rm sm}\asymp r_{\max}\asymp(k/n)^{1/d}$. (a) If $\kappa<d-4$, take $\theta_L=\Lambda^{2/(d-\kappa)}$, so that $\mathbb{E}W_2^2\lesssim\theta_L^2+n^{-1/2}\theta_{\rm sm}^{2-d/2}$; with $G=\theta_{\rm sm}/\theta_{\rm samp}>1$ the second term is θ_L^2 times $G^{-d(d-\kappa-4)/(2(d-\kappa))}\le 1$, and at the $\theta_{\rm sm}\asymp n^{-1/(2d-\kappa)}$ of (2), $\theta_L\asymp n^{-2/(2d-\kappa)}$, the residual (37) of W_1. If $\kappa=d-4$, take $\theta_L=\Lambda^{1/2}$: both sums are of order Λ, up to the factor $\log(\theta_{\rm sm}/\theta_L)\lesssim\log n$ below, and at $\theta_{\rm sm}\asymp n^{-1/(d+4)}$, $\Lambda\asymp n^{-4/(d+4)}$. (b) If $d-4<\kappa\le d-2$, let $\theta_L\to 0$: $\mathbb{E}W_2^2\lesssim n^{-1/2}\theta_{\rm sm}^{2-d/2}$, which at $k=\lceil n^{4/(d+4)}\rceil$, $\theta_{\rm sm}\asymp n^{-1/(d+4)}$, is $n^{-4/(d+4)}$. In both cases $Cr_{\max}^2$ is of the same order as the fluctuation, $Dp_1^{1/2}$ is of lower order for $n\ge n_0$, and n_0 also makes $r_{\max}\le r_1$; Minkowski's inequality over the four terms concludes, and $K\le d-3$ is $\kappa\le d-4$. The exponent $1-d/4$ of $\theta_{\rm sm}$ in (b) is free of κ: between $d-4$ and $d-2$ further directions do not help the quadratic chaining, and above $d-2$ the $\dot H^{-1}$ route of (i) takes over. The exact dyadic sums reproduce the three regimes. □

Two draws from one base. Lemma E.3 bounds $\Phi_x^\circ(Q)^2$ by the average of $\rho_{x,\sigma}(Q)^2$ over partner sets, and so it keeps the exponent of a single simplex. Two independent draws from Φ_x° share a partner only with probability of order $1/k$. Otherwise they lie in two simplices through the base which together span up to $2(K-1)$ directions, and they are close only if both lie close to the base or the two simplices nearly meet. The next lemma bounds their probability of being close directly. Write

$$\kappa_j\ :=\ \min\{2(K-1)-j,\ d-1\},\qquad j=0,\dots,K-1, \tag{63}$$

so that $\kappa_0=\min\{2(K-1),d-1\}$ and $\kappa_{K-1}=\kappa$.

Lemma E.9 (Two draws from one base)**.** *In (G2) with $d\ge 3$, let $z\sim\mu_0$ be independent of $\mathcal{D}_A$, and let Y,Y' be two independent draws from Φ_z° given $(z,\mathcal{D}_A)$, that is, with independent uniform partner sets σ,σ' and independent flat Dirichlet weights. Then for all $t>0$*

$$\mathbb{P}\big(\|Y-Y'\|\le t\big)\ \le\ C''_{K,d}\sum_{j=0}^{K-1}k^{-j}\Big(\frac{t}{h_\star}\wedge 1\Big)^{\kappa_j}, \tag{64}$$

with $C''_{K,d}$ depending on K, d and c_f only. Consequently

$$\mathbb{E}\sum_{Q\in\mathcal{Q}_\ell}\mathrm{Var}\left(\widehat{\nu}_X^\circ(Q)\mid\mathcal{D}_A\right)\ \le\ \frac{C''_{K,d}}{n_B}\sum_{j=0}^{K-1}k^{-j}\Big(\frac{\sqrt{d}D\theta_\ell}{h_\star}\wedge 1\Big)^{\kappa_j}. \tag{65}$$

Proof. The overlap. Let $J:=|\sigma\cap\sigma'|$. A fixed set of j members of σ lies in σ' with probability $\binom{k-j}{K-1-j}/\binom{k}{K-1}=\prod_{i<j}\frac{K-1-i}{k-i}\le\big(\frac{K-1}{k}\big)^j$, so $\mathbb{P}(J\ge j)\le\binom{K-1}{j}\big(\frac{K-1}{k}\big)^j$. It therefore suffices to show $\mathbb{P}(\|Y-Y'\|\le t\mid J=j)\le C(t/h_\star\wedge 1)^{\kappa_j}$ for each j.

Reduction to a small ball. A shared candidate that is a supplementary draw of the truncation is drawn afresh for each synthetic point, so it gives two independent partners; this only lowers the effective overlap, and we count such a candidate as unshared. Given $J=j$, the union $\sigma\cup\sigma'$ then has $q:=2(K-1)-j$ members. With $T_p:=p-z$,

$$Y-Y'\ =\ \sum_{p\in\sigma\cap\sigma'}(U_p-U'_p)\,T_p+\sum_{p\in\sigma\setminus\sigma'}U_pT_p-\sum_{p\in\sigma'\setminus\sigma}U'_pT_p\ =\ Mw,$$

where M is the $d\times q$ matrix of the offsets T_p, $p\in\sigma\cup\sigma'$. Suppose first that $q\le d-1$. The weights of σ and of σ' are independent, each with density $(K-1)!$ on its simplex, and integrating out U'_p for each of the j shared partners, over a fibre of length at most one, shows that w has density at most $((K-1)!)^2$ on $\mathbb{R}^q$. As in Step 1 of Proposition C.4, the preimage of a ball of radius t is an ellipsoid, so

$$\mathbb{P}\big(\|Y-Y'\|\le t\mid M\big)\ \le\ 1\wedge\frac{((K-1)!)^2\,\omega_q\,t^q}{\prod_{i\le q}\sigma_i(M)}. \tag{66}$$

If $q\ge d$, aggregate the weights as in Lemma E.2. Choose $A\subset\sigma$ and $A'\subset\sigma'$ with $|A\cup A'|=d-1$, and write $Y=BY_A+(1-B)Y_{\rm rest}$, where Y_A is a flat Dirichlet combination of $\{z\}\cup A$, independent of $B\sim\mathrm{Beta}(|A|+1,K-1-|A|)$ and of $Y_{\rm rest}$ ($B=1$ if $A=\sigma$); write Y' in the same way with A' and B'. Conditionally on B, B', $Y_{\rm rest}$ and $Y'_{\rm rest}$, the difference $Y-Y'$ is a constant plus $M'w'$, where M' is the matrix of the $d-1$ offsets of $A\cup A'$. The weights of Y_A are scaled by B and those

of $Y'_{A'}$ by B', and for a shared partner the difference loses one factor B' to a fibre of length at most B', so w' has density at most $|A|!\,|A'|!\,B^{-|A|}B'^{-|A'\setminus A|}$. The moments $\mathbb{E}B^{-|A|}$ and $\mathbb{E}B'^{-|A'|}$ are finite, because the first parameter of each Beta distribution exceeds the power, and $1 \wedge (\cdot)$ is concave, so (66) holds with q replaced by $d-1=\kappa_j$, M by M' and a constant depending on K and d.

Peeling. Condition on z and on $\rho := \min\{R_{k+1}(z), r_{\max}\}$. The candidates are then i.i.d. $\mu_0|_{B(z,\rho)}$, and σ, σ' are chosen independently of their positions, so the κ_j distinct partners behind M or M' are i.i.d. $\mu_0|_{B(z,\rho)}$. Since $\kappa_j \le d-1$, Step 3 of Proposition C.4 gives $\mathbb{E}[\rho^{\kappa_j}/\prod_i \sigma_i \mid z, \rho] \le \Lambda_{\kappa_j+1,d}$, and Step 4, as in Lemma E.2, gives $\mathbb{E}[(h_\star/\rho)^{\kappa_j}] \le c$. Multiplying proves (64) for $t < h_\star$; for $t \ge h_\star$ it is trivial.

The variance. Given $\mathcal{D}_A$, $\mathrm{Var}(\widehat{\nu}^\circ_X(Q) \mid \mathcal{D}_A) \le n_B^{-1}\mathbb{E}[\Phi^\circ_X(Q)^2 \mid \mathcal{D}_A]$, and $\sum_Q \Phi^\circ_x(Q)^2$ is the probability that two independent draws from Φ°_x fall in one cell, which is at most $\mathbb{P}(\|Y-Y'\| \le \sqrt{d}s \mid x, \mathcal{D}_A)$. Averaging over X and $\mathcal{D}_A$ and applying (64) at $t=\sqrt{d}s$ gives (65). □

The pool term would need the same bound, which we state as a hypothesis.

Assumption 3 (Two-point (D1) for the pool)**.** In (G2) there is C_{P}, depending on K, d and c_f only, such that for all n_A and k and every level ℓ

$$\sum_{Q\in\mathcal{Q}_\ell} \mathrm{Var}\left(\bar{\nu}_A(Q)\right) \;\le\; \frac{C_{\mathrm{P}}}{n_A}\sum_{j=0}^{K-1} k^{-j}\Big(\frac{\sqrt{d}D\theta_\ell}{h_\star}\wedge 1\Big)^{\kappa_j}. \tag{67}$$

Assumption 3 is (65) with the pool in place of the bases. Lemma E.4 proves a weaker form, with the single term $(\sqrt{d}D\theta_\ell/h_\star \wedge 1)^\kappa$ at weight one (59). That proof reduces the variance, through the Efron–Stein inequality, to $\sum_Q \sigma_r(Q)^2$ for at most six anchored measures σ_r per replaced pool point, and bounds each by $\sigma_r(\mathbb{R}^d)\sup_y \sigma_r(B(y,\sqrt{d}s))$, which sees one simplex at a time. Since $\sum_Q \sigma_r(Q)^2 = \sigma_r(\mathbb{R}^d)^2\,\mathbb{P}(Y, Y' \text{ in one cell})$ for two independent draws Y, Y' from $\sigma_r/\sigma_r(\mathbb{R}^d)$, and $\sigma_r(\mathbb{R}^d) \le 16(K-1)c_f/n_A$, Assumption 3 follows from the *anchored two-point bound*: such Y, Y' satisfy (64) on average over the anchor and the pool. This is the form in which we test it. The two draws come from two independent bases that both have the replaced point among their candidates. For the measures anchored at the leaving or the entering point, that point is the anchor of both. For the other two, the anchor is the neighbour of each base that enters or leaves in exchange, which is the $(k+1)$-st neighbour of the base in the pool that contains the replaced point, and it differs from base to base. The proof of Lemma E.9 does not carry over, for two reasons. The other partners of the two bases are candidates of overlapping neighbourhoods, so they are not conditionally i.i.d. given the bases. And for the latter two measures the anchor moves with the base.

We simulated both kinds of anchored measure on a uniform pool of $n = 20000$ points on $\mathbb{T}^d$. The bases are drawn from μ_0 and grouped by the replaced point, which is one of their k nearest neighbours. Over $t/h \in [\frac{1}{16}, \frac{1}{4}]$, where $h = (k/(n\omega_d))^{1/d}$ is the k-NN radius, we fit the log-log slope of $\mathbb{P}(\|Y-Y'\| \le t)$. For the measures anchored at the replaced point and $k = 24$, it is 1.88, 1.97, 2.00 and 2.00 for pairs at $d = 3, 4, 5, 6$, against the two-point exponent $\kappa_0 = 2$ and the single-simplex exponent 1. At $K = 3$ it is 3.54 and 3.67 for $d = 5, 6$, against $\kappa_0 = 4$ and 2. The shortfall at $K = 3$ is the term with one shared partner, of weight k^{-1} and exponent 3, and it shrinks as k grows: at $k = 96$ the slopes are 3.75 and 3.84, and 1.98 for pairs at $d = 4$. For the other two measures, whose anchors differ from base to base, the probabilities are 16 to 430 times smaller at every $t \le h/4$, with slopes 2.20–2.82 for pairs and 3.99 and 3.81 at $K = 3$, $d = 5, 6$ (3.41, 4.60 and 4.25 at $k = 96$). With a fixed radius, where the bases of an anchor are uniform on a ball around it, the slopes are 1.87 and 2.09 for pairs at $d = 3, 5$, and 3.7 and 3.9 at $K = 3$, $d = 5, 6$. For the bases, where Lemma E.9 is proved, the same fit gives 1.59–1.62 for pairs at $d = 3, \dots, 6$ and 2.94 and 2.99 at $K = 3$, $d = 5, 6$, against 0.86–0.91 and 1.65–1.70 for two points of one partner set. There the terms with a shared partner, $k^{-1}(t/h_\star)$ for pairs and $k^{-1}(t/h_\star)^3$ at $K = 3$, are visible at $k = 24$ and fade as k grows, as (64) says: at $k = 96$ the slopes are 1.86 for pairs at $d = 4$, and 3.40 and 3.49 at $K = 3$.

Proposition E.10 (The arity exponent under Assumption 3)**.** *In (G2) with $d \ge 3$, let Assumption 3 hold, put $\bar{\kappa}_2 := \min\{2(K-1), d-2\}$ and $k = \lceil n^{(d-\bar{\kappa}_2)/(2d-\bar{\kappa}_2)}\rceil$. There are C and n_0, depending only on d, K, Ω, f and C_{P}, such that for $n \ge n_0$*

$$\mathbb{E}\, W_1(\widehat{\nu}_X, \mu_0) \;\le\; C\, n^{-2/(2d-\bar{\kappa}_2)}\,\ell_n, \tag{68}$$

where $\ell_n = \log n$ if $2(K-1) = d-2$ and $\ell_n = 1$ otherwise. In particular, pairs attain $n^{-2/(d+2)}$ at $d = 3$ without the logarithm of Theorem A and at $d = 4$ up to a logarithm, and $n^{-1/(d-1)}$ for $d \geq 5$. The rate $n^{-2/(d+2)}$ holds for $K > d/2$, and at $K = d/2$ up to the logarithm, instead of from $K = d-1$ on.

Proof. By (56) only the two fluctuation terms change. Lemma E.9 and Assumption 3 give both of them the cell-mass variance $\frac{C}{n}\sum_j k^{-j}(\theta_\ell/\theta_{\mathrm{sm}} \wedge 1)^{\kappa_j}$. Run Step 1 of Section B with it and use $\sqrt{\sum_j a_j} \leq \sum_j \sqrt{a_j}$. The level sum then splits into K sums of the form (34), the j-th with $S_j := Ck^{-j}/n$ and exponent κ_j, which share the finest-cell term $\sqrt{d}D\theta_L$. Choose for each j a near-minimiser $\theta^{(j)}$ of (27), and for θ_L a dyadic level between the largest of them and twice it. This can only lower the other sums, and $\theta_L \leq 2\sum_j \theta^{(j)}$, so the fluctuation is at most $c_d D \sum_j \big(\sqrt{S_j}\,\Phi_d(\theta_\star) + R_{d,\kappa_j}(\theta_\star)\big)$, with S_j in place of S in the j-th residual and $\theta_\star = \theta_{\mathrm{sm}} \asymp (k/n)^{1/d}$ for every j, since $S_j \leq S_0$. By Step 4 of Section B, with $S_j^{1/d}$ in place of θ_{samp}, the j-th term is of order $n^{-1/d}k^{-e_j}$, where

$$e_j \;=\; \frac{jd+\kappa_j}{d(d-\kappa_j)} \quad \text{if } \kappa_j \leq d-2, \qquad e_j \;=\; \frac{j}{2} + \frac{d-2}{2d} \quad \text{if } \kappa_j \geq d-2,$$

the two agreeing at $\kappa_j = d-2$, where a factor $\log k$ appears. No later term exceeds the first, because $e_j > e_0$ for $j \geq 1$, and the strict inequality absorbs the logarithm. If $\kappa_0 < d-2$, then $\kappa_j = \kappa_0 - j$, and $e_j > e_0$ reduces to $(d-1)(d-\kappa_0) > \kappa_0$. If $\kappa_0 \geq d-2$, then $e_j > e_0$ is clear when $\kappa_j \geq d-2$, and otherwise $\kappa_j \geq d-2-j$ gives $2(jd+\kappa_j) - (d-2)(d-\kappa_j) \geq jd > 0$. The first term is Step 4 with the exponent κ_0, which exceeds $d-2$ exactly when $2(K-1) > d-2$. As in the proof of Theorem A′(iii), with $\bar\kappa_2$ in place of $\bar\kappa$, the fluctuation is $O(n^{-2/(2d-\bar\kappa_2)}\ell_n)$ at the stated k, the bias is $O(h_\star^2) = O(n^{-2/(2d-\bar\kappa_2)})$, and the truncation terms are of lower order. Exact dyadic sums of this bound, with the hypergeometric weights of the overlap in place of k^{-j}, reproduce the exponent $2/(2d-\bar\kappa_2)$ for $d = 4, \ldots, 10$ and $K = 2, 3, 4$, the fitted value falling short only by the logarithm where $2(K-1) = d-2$. □

F THE W_2 BOUNDS

In W_2 the transportation profile is

$$\varphi_d(S) := S^{1/4}\ (1 \leq d \leq 3), \qquad S^{1/4}\sqrt{1+\log(1/S)}\ (d=4), \qquad S^{1/d}\ (d \geq 5), \tag{69}$$

with the explicit constants $\widetilde{C}_d = \sqrt{d}\,(1+4q/(1-q))^{1/2}$ for $1 \leq d \leq 3$, $\widetilde{C}_4 = 2\sqrt{5}$ and $\widetilde{C}_d = \sqrt{d}\,(2+2\sqrt{q/(q-1)})$ for $d \geq 5$, where $q := 2^{d/2-2}$, multiplied by $\sqrt{b_K}$ for $\widehat{\nu}_X$ as in Section A.4 Step 7. The deviation proxy is $V_2 := \max_i \alpha_i w_i$; since $S \leq w_{\max} \leq \sqrt{S}$, $V_2 \leq \alpha w_{\max} \leq \alpha\sqrt{S}$ when $\alpha_i \leq \alpha$.

On the torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$ we read Assumption 1 with $\Omega = \mathbb{T}^d$ and periodic distances, combine each tuple in the chart centred at its base (at its first point in (G1)), with offsets taken as shortest representatives, and write $\|\sigma\|_{\dot H^{-1}}^2 := \sum_{\xi\neq 0} |\hat\sigma(\xi)|^2/(4\pi^2|\xi|^2)$ for a finite signed measure σ, with $\hat\sigma(\xi) = \int e^{-2\pi i \xi\cdot y}\,\sigma(dy)$. The following is the W_2 counterpart of Theorem A′, for the same generators.

Theorem F.1 (Convex augmentation in W_2)**.** *Let $\widehat{\nu}_X$ be the synthetic distribution of (G1) or (G2) in Theorem A′, under the hypotheses there, let $\bar\nu$, S, $h_\star$ and κ be as there, and put $\beta_2 := W_2(\bar\nu, \mu_0)$. Then:*

(i) fluctuation: $\mathbb{E}\,W_2(\widehat{\nu}_X, \bar\nu) \leq \widetilde{C}_{d,K} D\,\varphi_d(S)$. *If moreover $d \geq 3$, μ_0 lives on $\mathbb{T}^d$, and either (G1) holds with (46) and $\kappa > d-2$, or (G2) holds with $K \geq d$ and $r_{\max} \leq r_1$, a constant depending on d, K and f, then*

$$\mathbb{E}\,W_2(\widehat{\nu}_X, \bar\nu) \;\leq\; c_{d,K}\,\bar f^{-1/2}\sqrt{S}\,h_\star^{1-d/2}, \tag{70}$$

where $\bar f = f_{\min}$ in (G2), and in (G1) $\bar f > 0$ is a lower bound on the density of $\bar\nu$, assumed to exist; in (G2) both bounds carry an additional $2De^{-k/4}$;

(ii) deviation: *for every* $\delta \in (0,1)$*, with probability at least* $1-\delta-p_n$*,*

$$\big|W_2(\widehat{\nu}_X,\mu_0) - \mathbb{E}W_2(\widehat{\nu}_X,\mu_0)\big| \;\le\; \Delta_\delta^{(2)}, \tag{71}$$

where $p_n = 0$ *in (G1),* $p_n = ne^{-k/2}$ *in (G2), and*

$$\Delta_\delta^{(2)} = \sqrt{2V_2\log(2/\delta)} \text{ in (G1)}, \qquad \Delta_\delta^{(2)} = 12\sqrt{(K-1)c_f}\,D\Big(\frac{\log(2/\delta)}{n}\Big)^{1/4} \text{ in (G2)};$$

(iii) tuned rate: *in (G2) on* $\mathbb{T}^d$ *with* $d \ge 3$*,* $K \ge d$ *and* k *from (2), which here reads* $k = \lceil n^{2/(d+2)}\rceil$*,* $\beta_2 = O(h_\star^2)$*, and there are* C *and* n_0*, depending only on* d*,* K *and* f*, such that* $\big(\mathbb{E}\,W_2^2(\widehat{\nu}_X,\mu_0)\big)^{1/2} \le C\,n^{-2/(d+2)}$ *for* $n \ge n_0$*.*

In (G1), $\widetilde{C}_{d,K} = \sqrt{b_K}\,\widetilde{C}_d$ *and* $c_{d,K}$ *depends on the constants of (46); in (G2) both constants depend on* c_f *as well. The (G1) part of (ii) assumes that the coordinates satisfy log-Sobolev inequalities with constants* α_i*, even on a bounded support (Remark A.9).*

Remark F.2 (What changes from Theorem A′). The three parts match those of Theorem A′; the differences are where W_2 asks for more. The refined fluctuation (70) comes from Peyre's inequality $W_2(\mu,\nu) \le 2\|\mu-\nu\|_{\dot H^{-1}(\nu)}$ rather than from chaining, so it needs a reference density bounded below, no boundary, and $\kappa > d-2$, the range where the Newtonian energy of a κ-dimensional blur is finite (Remark C.9): for SMOTE this is $K \ge d$, where Theorem A′ takes $K \ge 2$. Below $K = d$ the quadratic chaining of Steps 2–3 still runs on (D1): it recovers the rate of Theorem A for $K \le d-3$ and gives $n^{-2/(d+4)}$ for $K = d-2, d-1$ (Corollary F.5). With a boundary the bias itself is larger. Mixing depletes a collar of width $h_\star$ by a mass of order $h_\star$ moved a distance of order $h_\star$, which costs $h_\star^2$ in W_1 but $h_\star^{3/2}$ in W_2. For $d = 1$, $K = 2$ and μ_0 uniform on $[0,1]$ the fitted slope of β_2 over $h_\star \in [0.025, 0.2]$ is 1.49; on the circle with $f = 1 + 0.6\cos 2\pi x$, $\beta_2/h_\star^2 = 0.29$ and 0.31 at $h_\star = 0.1$ and 0.05, against the value 0.312 that linearisation predicts. This is why (iii) is stated on the torus; the tuned W_2 rate on a domain with boundary is open. Finally, the (G2) deviation is of order $n^{-1/4}$, not $n^{-1/2}$: bounded differences control W_2^2, whose changes are of order $1/n$, and Lemma F.6, which gives $\sqrt{V_2}$ in (G1), holds the tuples fixed, while k-NN partners move with the data. This deviation is of lower order than the rate $n^{-2/(d+2)}$ only for $d \ge 7$; at $d = 6$ the two match up to the logarithm. Under balanced reuse in (G1) it is of order $n^{-1/2}$, as in W_1 (Corollary F.3).

Corollary F.3 (W_2 under balanced reuse)**.** *In (G1) with* $w_i = 1/n$ *and* $\alpha_i \le \alpha$*, with probability at least* $1-\delta$*,* $W_2(\widehat{\nu}_X,\mu_0) \le \beta_2 + \widetilde{C}_{d,K} D\varphi_d(1/n) + \sqrt{2\alpha\log(1/\delta)/n}$*, and under the torus hypotheses of Theorem F.1(i) the middle term is* $c_{d,K}\bar f^{-1/2} n^{-1/2} h_\star^{1-d/2}$*. The deviation is that of an i.i.d. sample of size* n*: in* W_2*, as in* W_1*, reuse costs only through the fluctuation.*

Corollary F.4 (Tuned SMOTE beats its sample in W_2)**.** *In the setting of Theorem F.1(iii), with* $\hat\mu_n$ *the empirical measure of the sample,* $\big(\mathbb{E}W_2^2(\widehat{\nu}_X,\mu_0)\big)^{1/2}/\mathbb{E}W_2(\hat\mu_n,\mu_0) \le C' n^{-(d-2)/(d(d+2))} \to 0$*, the gain (4) at* $K \ge d$*.*

Corollary F.5 (Tuned W_2 below arity d)**.** *In (G2) on* $\mathbb{T}^d$*, under the hypotheses of Theorem F.1(iii) except that* $d \ge 5$ *and* $2 \le K \le d-1$*, there are* C *and* n_0*, depending only on* d*,* K *and* f*, such that for* $n \ge n_0$*: (a) if* $K \le d-3$*, with* k *from (2),* $\big(\mathbb{E}W_2^2(\widehat{\nu}_X,\mu_0)\big)^{1/2} \le Cn^{-2/(2d-\bar\kappa)}\ell'_n$*, the rate of Theorem A(i), where* $\ell'_n = \sqrt{\log n}$ *at* $K = d-3$ *and* $\ell'_n = 1$ *below; (b) if* $K \in \{d-2, d-1\}$*, with* $k = \lceil n^{4/(d+4)}\rceil$*,* $\big(\mathbb{E}W_2^2(\widehat{\nu}_X,\mu_0)\big)^{1/2} \le Cn^{-2/(d+4)}$*, which beats* $n^{-1/d}$ *but not the* $n^{-2/(2d-\bar\kappa)}$ *of* W_1*. The same holds for the* ε*-ball rule. Pairs fall under (a) from* $d = 5$ *on, with* $\ell'_n = \sqrt{\log n}$ *at* $d = 5$*.*

Proof of Theorem F.1 for (G1). (i) The first bound is Steps 1–6 below. The second is Proposition C.8 with $f_{\min} = \bar f$, whose $\theta_\star^{1-d/2}$ is $(h_\star/C_Q)^{1-d/2}$ at $D = 1$. (ii) Step 7 uses of $\bar\nu$ only that it does not depend on X, so Lemma F.6 holds for $X \mapsto W_2(\widehat{\nu}_X,\mu_0)$, and Step 8, applied to this map and to its negative, gives $\mathbb{P}(|W_2(\widehat{\nu}_X,\mu_0) - \mathbb{E}W_2(\widehat{\nu}_X,\mu_0)| \ge t) \le 2e^{-t^2/(2V_2)}$. Part (iii) does not concern (G1). The proof for (G2) is in Section E. □

Proofs of the corollaries. *Corollary F.3.* With $w_i = 1/n$, $w_{\max} = 1/n$ and $V_2 \le \alpha/n$; insert in (77) of Step 8, whose deviation is around the mean whatever bounds it, and bound the mean by

Theorem F.1(i). *Corollary F.4.* $W_2 \geq W_1$, and the lower bound $W_1(\hat{\mu}_n, \mu_0) \geq c\, n^{-1/d}$ in the proof of (4) holds on $\mathbb{T}^d$ as well; divide Theorem F.1(iii) by it and use $\frac{2}{d+2} - \frac{1}{d} = \frac{d-2}{d(d+2)}$. Corollary F.5 is proved at the end of Section E. □

Lemma F.6 (Transport-weighted gradient)**.** *Let* $H_2(X) := W_2(\widehat{\nu}_X, \bar{\nu})$ *and, for* $\alpha_i > 0$*, let* $d_\alpha(X, X')^2 := \sum_i \|X_i - X'_i\|_2^2/\alpha_i$*. Then for all* X, X'*,*

$$|H_2(X) - H_2(X')| \;\leq\; \sqrt{V_2}\, d_\alpha(X, X'), \qquad V_2 = \max_{1\leq i\leq n} \alpha_i w_i, \tag{72}$$

equivalently $\sum_i \alpha_i \|\nabla_{X_i} H_2\|_2^2 \leq V_2$ *wherever* H_2 *is differentiable. In particular* $\mathrm{Lip}(H_2) \leq \sqrt{w_{\max}} \leq S^{1/4}$*. Coordinates with* $\alpha_i = 0$ *are a.s. constant and are held fixed.*

Step 1: the fluctuation–bias split. By the triangle inequality for W_2, $W_2(\widehat{\nu}_X, \mu_0) \leq W_2(\widehat{\nu}_X, \bar{\nu}) + W_2(\bar{\nu}, \mu_0) = H_2(X) + \beta_2$, with β_2 deterministic. No dual identity is needed for this: unlike the W_1 case, where (17) additionally identifies the fluctuation exactly, here the triangle inequality alone already removes the interpolation term $\eta_{\mathrm{int},2} = \mathbb{E}W_2(\widehat{\nu}_X, \mu_w)$ that a μ_w-based route would carry. It remains to bound $\mathbb{E}H_2$ and to concentrate H_2. Steps 2–6 bound $\mathbb{E}W_2(\mu_w, \mu_0)$; the same estimates apply verbatim to $\mathbb{E}W_2(\widehat{\nu}_X, \bar{\nu})$ with $\bar{\nu}$ in place of μ_0 and $\widehat{\nu}_X$ in place of μ_w, since the dyadic argument uses only $\mathbb{E}\Delta_Q = 0$ and the variance bound of Section A.4 Step 7.

Step 2: quadratic dyadic estimate. With the notation of Section A.4 Step 3, for every integer $L \geq 1$,

$$W_2^2(\mu_w, \mu_0) \;\leq\; dD^2 2^{-2L} + dD^2 \sum_{\ell=1}^{L} 2^{-2(\ell-1)} A_\ell, \tag{73}$$

by the same recursive matching, now with squared costs: the within-cell quadratic cost at the finest level is at most $dD^2 2^{-2L}$ and the squared parent diameter at level ℓ is at most $dD^2 2^{-2(\ell-1)}$.

Step 3: expected discrepancy. Exactly as in (23), $\mathbb{E}A_\ell \leq \sqrt{S} 2^{d\ell/2}$, so

$$\mathbb{E}W_2^2(\mu_w, \mu_0) \;\leq\; dD^2 2^{-2L} + 4dD^2\sqrt{S} \sum_{\ell=1}^{L} 2^{\ell(d/2-2)}, \qquad \mathbb{E}W_2(\mu_w, \mu_0) \leq \sqrt{\mathbb{E}W_2^2(\mu_w, \mu_0)}.$$

Step 4: optimisation for $1 \leq d \leq 3$**.** Here $d/2 - 2 < 0$, so $\sum_{\ell\leq L} 2^{\ell(d/2-2)} \leq \frac{2^{d/2-2}}{1-2^{d/2-2}}$. Keeping the finest-cell term and choosing $L = \lceil \frac{1}{4}\log_2(1/S) \rceil$ gives $\mathbb{E}W_2^2(\mu_w, \mu_0) \leq dD^2 S^{1/2}(1 + \frac{4\cdot 2^{d/2-2}}{1-2^{d/2-2}})$, hence $\mathbb{E}W_2(\mu_w, \mu_0) \leq \sqrt{d} D (1 + \frac{4\cdot 2^{d/2-2}}{1-2^{d/2-2}})^{1/2} S^{1/4}$.

Step 5: optimisation for $d = 4$**.** Here $\sum_{\ell\leq L} 2^{\ell(d/2-2)} = L$. Choosing $L = \lceil \frac{1}{4}\log_2 \frac{1}{S} \rceil$ gives $2^{-2L} \leq S^{1/2}$ and $L \leq 1 + \frac{1}{4}\log_2 \frac{1}{S}$, so $\mathbb{E}W_2^2(\mu_w, \mu_0) \leq 4D^2 S^{1/2} + 16 D^2 S^{1/2}(1 + \frac{1}{4}\log_2 \frac{1}{S}) = D^2 S^{1/2}\big(20 + 4\log_2 \frac{1}{S}\big) \leq 20 D^2 S^{1/2}(1 + \log \frac{1}{S})$, the last step because $4/\log 2 < 20$ (the constant 16 would fail for $S > 0.68$), and therefore $\mathbb{E}W_2(\mu_w, \mu_0) \leq 2\sqrt{5}\, D S^{1/4} \sqrt{1 + \log(1/S)}$, i.e. $\widetilde{C}_4 = 2\sqrt{5}$.

Step 6: optimisation for $d \geq 5$**.** Let $r := 2^{d/2-2} > 1$ and $L = \lfloor \frac{1}{d}\log_2 \frac{1}{S} \rfloor$, so $2^{-L} \leq 2S^{1/d}$ and $\sqrt{S}\sum_{\ell\leq L} 2^{\ell(d/2-2)} \leq \frac{r}{r-1} S^{2/d}$. Then $\mathbb{E}W_2^2(\mu_w, \mu_0) \leq dD^2(4 + \frac{4r}{r-1}) S^{2/d}$ and, using $\sqrt{4 + \frac{4r}{r-1}} \leq 2 + \sqrt{\frac{4r}{r-1}}$, $\widetilde{C}_d = \sqrt{d}\,(2 + \sqrt{\frac{4\cdot 2^{d/2-2}}{2^{d/2-2}-1}})$.

Step 7: transport-weighted Lipschitz bound (Lemma F.6). Write $Y = Y_\alpha(X, u^\alpha)$ for a generic synthetic point and, for each i,

$$\sigma_i(\alpha, u) := \sum_{j:\, \alpha_j = i} u_j^\alpha, \qquad \text{so that} \qquad \sum_{i=1}^{n} \sigma_i \equiv 1, \qquad \mathbb{E}\sigma_i = w_i, \tag{74}$$

where $\mathbb{E}$ denotes the average $\frac{1}{n_\pi}\sum_\alpha \mathbb{E}_{u^\alpha}$; the identity $\mathbb{E}\sigma_i = \frac{1}{n_\pi}\sum_\alpha c_{\alpha,i}/K = r_i/(n_\pi K) = w_i$ is the computation of Section A.3 Step 1. No optimal map is used, since $\widehat{\nu}_X$ is singular for $K \le d$. Let γ be an optimal plan between $\widehat{\nu}_X$ and $\bar{\nu}$ and, by the gluing lemma, let π couple the generator randomness (α, u) with $Z \sim \bar{\nu}$ so that $(Y, Z) \sim \gamma$. Moving X to $X + \eta$ moves Y to $Y + \delta$ with $\delta := \sum_i \sigma_i \eta_i$, and $(Y + \delta, Z)$ is a coupling of $\widehat{\nu}_{X+\eta}$ with $\bar{\nu}$, so

$$H_2(X+\eta)^2 \;\le\; H_2(X)^2 + 2\,\mathbb{E}_\pi[(Y-Z)\cdot\delta] + \mathbb{E}_\pi\|\delta\|^2. \tag{75}$$

Put $c_i := \mathbb{E}_\pi[\sigma_i\|Y-Z\|^2]$, the share of the transport cost carried by X_i; because $\sum_i \sigma_i \equiv 1$, the shares obey the conservation law

$$\sum_{i=1}^{n} c_i \;=\; \mathbb{E}_\pi\|Y-Z\|^2 \;=\; H_2(X)^2. \tag{76}$$

Cauchy–Schwarz with the weight σ_i split as $\sqrt{\sigma_i}\cdot\sqrt{\sigma_i}$ gives $\mathbb{E}_\pi[\sigma_i\|Y-Z\|] \le \sqrt{w_i c_i}$, hence

$$\big|\mathbb{E}_\pi[(Y-Z)\cdot\delta]\big| \le \sum_i \|\eta_i\|\sqrt{w_i c_i} \le d_\alpha(\eta)\Big(\sum_i \alpha_i w_i c_i\Big)^{1/2} \le \sqrt{V_2}\, d_\alpha(\eta)\, H_2(X),$$

using (76) in the last step, and by Jensen with $\sum_i \sigma_i \equiv 1$, $\mathbb{E}_\pi\|\delta\|^2 \le \sum_i w_i\|\eta_i\|^2 \le V_2\, d_\alpha(\eta)^2$. Together, $H_2(X+\eta)^2 \le \big(H_2(X) + \sqrt{V_2}\, d_\alpha(\eta)\big)^2$. Exchanging X and $X+\eta$ gives (72) for every pair of points, $H_2(X) = 0$ included. Taking $\alpha_i \equiv 1$ gives $\mathrm{Lip}(H_2)^2 \le w_{\max} \le \sqrt{S}$.

Step 8: concentration and assembly. Mollify H_2 on $(\mathbb{R}^d)^n$: the result is smooth, still $\sqrt{V_2}$-Lipschitz in d_α, hence $\sum_i \alpha_i\|\nabla_{X_i}\cdot\|^2 \le V_2$ at every point, and it converges to H_2 uniformly. The log-Sobolev inequality is thus applied to smooth functions only, so no absolute continuity of the μ_i is needed (conditionally on $\mathcal{D}_A$ in Section G some coordinates are Dirac masses). By tensorisation and the Herbst argument (Section A.3 Steps 3–5), passing to the limit, $\mathbb{E}\exp(\lambda(H_2 - \mathbb{E}H_2)) \le \exp(\lambda^2 V_2/2)$ and $\mathbb{P}(H_2 \ge \mathbb{E}H_2 + t) \le \exp(-t^2/(2V_2))$, so with probability at least $1-\delta$, $H_2 \le \mathbb{E}H_2 + \sqrt{2V_2\log(1/\delta)}$. With Step 1 and the expectation bound of Steps 2–6, with probability at least $1-\delta$,

$$W_2(\widehat{\nu}_X, \mu_0) \;\le\; \beta_2 + \widetilde{C}_{d,K}\, D\, \varphi_d(S) + \sqrt{2V_2\log(1/\delta)}. \tag{77}$$

If $\alpha_i \le \alpha$ for all i then $V_2 \le \alpha w_{\max} \le \alpha\sqrt{S}$. □

Remark F.7 (Why the naive bound is vacuous, and what repairs it). Bounding $|H_2(X) - H_2(X')|$ by $W_2(\widehat{\nu}_X, \widehat{\nu}_{X'})$ and coupling with the same (α, u^α) on both sides gives, by convexity of $x \mapsto \|x\|_2^2$ and $\mathbb{E}u_j^\alpha = 1/K$, $W_2^2(\widehat{\nu}_X, \widehat{\nu}_{X'}) \le \sum_i w_i\|X_i - X_i'\|_2^2$, i.e. coordinatewise constants $\sqrt{w_i}$ and the proxy $\sum_i \alpha_i w_i \in [\min_i \alpha_i, \max_i \alpha_i]$. That quantity never decays, because $\sum_i w_i = 1$ is a constraint, not a small number. The corresponding deviation term $\sqrt{2\alpha\log(1/\delta)}$ therefore stays of constant order, and for a distribution of diameter D, where $\alpha \asymp D^2$, it is larger than the trivial bound $W_2 \le D$. Step 7 avoids this by never passing through a worst-case coordinatewise constant: each block is weighted by its own share c_i of the transport cost, and the conservation law (76) converts the sum into a maximum. The same device does not improve V_1, where the constants w_i already square to the summable S.

G SMOTE IN PRACTICE

Let G_ε have an edge (i,j) iff $\|X_i - X_j\| < \varepsilon$, let $N(i)$ be the neighbourhood of vertex i, in G_ε or in the k-NN graph, and $D_i := |N(i)|$ its degree.

Proposition G.1 (Overlap of the SMOTE sampling distribution)**.** *Let π draw a base b uniformly from the sample and complete it with $K-1$ distinct neighbours drawn uniformly from $N(b)$, and put $T_i := \sum_{j:\, i\in N(j)} 1/D_j$, so that $\frac{1}{n}\sum_i T_i = 1$. Then, as an identity,*

$$p_i = \frac{1 + (K-1)T_i}{Kn}, \qquad S_\pi = \frac{1}{n}\Big(1 + \Big(\frac{K-1}{K}\Big)^2 \mathrm{Var}(T)\Big), \quad \mathrm{Var}(T) := \tfrac{1}{n}\textstyle\sum_i (T_i - 1)^2. \tag{78}$$

The inflation over $1/n$ is thus a variance: measured, nS_π is 1.1–1.4 for a uniform base, a constant factor where a worst-case count would suggest $S \gg 1/n$.

Step 0: the draw, and why n_π drops out. With $\alpha^1, \dots, \alpha^{n_\pi}$ i.i.d. $\sim \pi$ the multiplicities are sums of n_π i.i.d. shares, $r_i = \sum_{t \le n_\pi} c_i^t$ with $\mathbb{E}c_i = Kp_i$ and $c_i \le K$, so $\mathbb{E}r_i = n_\pi K p_i$, $\mathrm{Var}(r_i) = n_\pi \mathrm{Var}(c_i) \le n_\pi K \mathbb{E}c_i = n_\pi K^2 p_i$ and

$$\mathbb{E}S = \sum_i \frac{(\mathbb{E}r_i)^2 + \mathrm{Var}(r_i)}{(n_\pi K)^2} = S_\pi + \frac{1}{n_\pi K^2} \sum_i \mathrm{Var}(c_i) \;\le\; S_\pi + \frac{1}{n_\pi}. \tag{79}$$

Replacing one draw changes w by at most $2/n_\pi$ in ℓ^1 and $1/n_\pi$ in ℓ^∞, so $|S(\alpha) - S(\alpha')| \le 4w_{\max}/n_\pi$ and bounded differences give $\mathbb{P}(|S - \mathbb{E}S| \ge t) \le 2\exp(-t^2 n_\pi / 8w_{\max}^2)$. Both statements are free of any assumption on π; in particular sampling beyond the number of distinct tuples neither helps nor hurts, and $S \ge 1/n$ holds for every realisation by Remark A.4.

Step 1: the uniform-base distribution. Let π draw b uniformly and then $K-1$ distinct neighbours uniformly from $N(b)$. Then $c_i = \mathbf{1}\{b = i\} + \#\{\text{partners} = i\}$, and since a partner slot lands on i only through a base j with $i \in N(j)$, $\mathbb{E}c_i = \frac{1}{n} + \frac{K-1}{n} T_i$ with $T_i = \sum_{j:\, i \in N(j)} 1/D_j$; summing gives $\frac{1}{n}\sum_i T_i = 1$ and $p_i = (1 + (K-1)T_i)/(Kn)$ as claimed. Hence

$$S_\pi = \frac{1}{K^2 n^2} \sum_i \big(1 + (K-1)T_i\big)^2 = \frac{1}{K^2 n}\Big[\big(1 + (K-1)\big)^2 + (K-1)^2\,\mathrm{Var}(T)\Big],$$

which is (78); the cross term is $2(K-1)n$ because T has mean one, and that is what makes the identity exact rather than an inequality. For a k-NN graph $D_j \equiv k$, so $T_i = d_i^{\mathrm{in}}/k$ and $\mathrm{Var}(T) = \mathrm{Var}(d^{\mathrm{in}})/k^2$, bounded by a constant depending only on d since $d_i^{\mathrm{in}} \le \gamma_d k$. □

A data-dependent index distribution, and the splitting repair. Everything above is stated for a π that does not depend on X. When the neighbourhood graph is built from the same sample, π and hence w are functions of X, and the tensorisation behind Proposition A.8 no longer applies. Splitting repairs the expectation bounds: partition the sample into $\mathcal{D}_A, \mathcal{D}_B$ of size $n/2$, build G_ε on $\mathcal{D}_A$ only, and complete each point of $\mathcal{D}_B$ with its ε-neighbours *in* $\mathcal{D}_A$. Conditionally on $\mathcal{D}_A$ the bases stay i.i.d. and distinct bases give independent components, so every variance bound behind an expectation bound holds with S the overlap of the bases, $1/n_B$ for uniform bases. The deviation bounds do not carry over verbatim. The partners of a base are a function of its position, piecewise constant with jumps where the neighbourhood changes, so moving X_b moves the tuple and Lemma A.6 and Lemma F.6, which hold the tuple fixed, do not apply: conditionally on $\mathcal{D}_A$ the index distribution is *not* fixed. Theorem A′(ii) replaces the log-Sobolev argument by bounded differences, which pay D/n_B per base and need no Lipschitz property (Section E). The bias side needs no repair at all: the base point is drawn uniformly, so coupling each synthetic point to its own base gives a second marginal equal to the unweighted empirical measure, whose expectation is μ_0 exactly, and a data-adaptive π contributes no selection bias to the first-order bound $\beta_1 \le \varepsilon$.

The second-order bias of Proposition 4.1 is a different matter. It is computed for partners drawn from $\mu_0|_{B(Z_1, h_\star)}$, whereas after splitting they are drawn from $\mathcal{D}_A$, so conditionally on $\mathcal{D}_A$ the synthetic mean is $\bar\nu_A := \mathbb{E}[\widehat\nu_X \mid \mathcal{D}_A]$, the base point being integrated against μ_0, and $W_1(\widehat\nu_X, \mu_0) \le W_1(\widehat\nu_X, \bar\nu_A) + W_1(\bar\nu_A, \bar\nu) + \beta_1$. The first term is Theorem B.3 conditionally on $\mathcal{D}_A$ and the last is Proposition 4.1. For both neighbourhood rules and every arity $2 \le K \le d$, Lemma E.4 bounds the middle one by the fluctuation's order with the exponent κ, which is all Theorem A needs.

Lemma G.2 (k-nearest-neighbour radius)**.** *Suppose $\mu_0(B(x,r)) \ge c_0 r^d$ for every $x \in \operatorname{supp}\mu_0$ and $0 < r \le D$, and let $R_k(X_i)$ be the distance from X_i to its k-th nearest neighbour among the other $n-1$ points. Then with probability at least $1-\delta$,*

$$\max_{i \le n} R_k(X_i) \;\le\; \Big(\frac{2k + 8\log(n/\delta)}{c_0\,(n-1)}\Big)^{1/d} \;\asymp\; \Big(\frac{k + \log(n/\delta)}{n}\Big)^{1/d}. \tag{80}$$

Proof. Condition on $X_i = x$ and put $q := \mu_0(B(x,r)) \ge c_0 r^d$. The number of the other $n-1$ points in $B(x,r)$ is $\mathrm{Bin}(n-1, q)$, and $R_k(X_i) > r$ forces it below k. Choose r with $(n-1)c_0 r^d = 2k + 8\log(n/\delta)$, so that $(n-1)q \ge 2k$ and hence $k \le (n-1)q/2$; the multiplicative Chernoff bound $\mathbb{P}(\mathrm{Bin}(N,q) \le Nq/2) \le e^{-Nq/8}$ then gives $\mathbb{P}(R_k(X_i) > r) \le e^{-k/4}\,\delta/n \le \delta/n$, and a union bound over i finishes the proof. □

TUNING THE LOCALITY SCALE

Everything above fixes the generator and reads off what it costs. Reversing the question, and choosing $h_\star$ to minimise the *total* error, turns (1) into an ordinary bias–variance problem, and its solution is a convex-mixing scheme that provably outruns the empirical measure. Theorem A is the rigorous form of this for k-NN SMOTE, with the split, the pool term and the k-NN bias accounted for (Section E); here we optimise the bound itself.

Let $d \geq 3$, let μ_0 have a C^2 density with $\nabla f \not\equiv 0$ on a bounded domain, let reuse be balanced ($S = 1/n$), and let the generator be $h_\star$-local with arity K and $\kappa = \min(K-1, d-1)$, $h_\star$ being the partner radius of Proposition 4.1; its spread $2h_\star$, which is what enters F_κ, changes F_κ by a constant factor only. Adding Proposition 4.1 to Theorem B.3 and Proposition B.4,

$$\begin{aligned} \mathbb{E}W_1(\widehat{\nu}_X, \mu_0) \;&\leq\; \underbrace{C_\beta h_\star^2}_{\beta_1} + F_\kappa(h_\star), \\ F_\kappa(h) = c_{d,\kappa} D\, n^{-1/(d-\bar\kappa)} \Big(\frac{C_Q D}{h}\Big)^{\bar\kappa/(d-\bar\kappa)}, \qquad &\bar\kappa := \min(\kappa, d-2), \end{aligned} \tag{81}$$

with $C_\beta = \frac{K-1}{(d+2)(K+1)}\|\nabla f\|_{L^1} + C_\partial$, and an extra factor $\log(h/(C_Q D n^{-1/d}))$ in F_κ at $\kappa = d-2$. For $\kappa \geq d-2$ this is $c_d D n^{-1/2}(C_Q D/h)^{d/2-1}$, case (i) of Proposition B.4; for $\kappa < d-2$ it is (37) with $\theta_{\mathrm{samp}} = n^{-1/d}$.

Proposition G.3 (Optimal locality, and a rate below the i.i.d. one)**.** *The right-hand side of (81) is minimised at $h_\star^\star/D \asymp n^{-1/(2d-\bar\kappa)}$, which is $n^{-1/(d+2)}$ for $\kappa \geq d-2$. The neighbourhood follows $k \asymp n h_\star^d$, that is*

$$k \;\asymp\; n^{(d-\bar\kappa)/(2d-\bar\kappa)}, \tag{82}$$

the exponent being capped because $h_\star^\star$ stops moving once κ passes $d-2$. At that choice both terms are of the same order, so

$$\mathbb{E}W_1(\widehat{\nu}_X, \mu_0) \;\lesssim\; n^{-\frac{2}{2d-\bar\kappa}}, \qquad \textit{that is } n^{-\frac{2}{d+2}} \textit{ for } \kappa \geq d-2, \tag{83}$$

with an extra $(\log n)^{4/(d+2)}$ at $\kappa = d-2$. The exponent is strictly increasing in $\bar\kappa$, equals $1/d$ exactly at $\kappa = 0$, and exceeds $1/d$ for every $\kappa \geq 1$.

Proof. Differentiate (81): with $q := \bar\kappa/(d-\bar\kappa)$ and $\Lambda \propto n^{-1/(d-\bar\kappa)}$, $2C_\beta h = q\Lambda h^{-q-1}$, so $h^{q+2} \propto \Lambda$, i.e. $h^{(2d-\bar\kappa)/(d-\bar\kappa)} \propto n^{-1/(d-\bar\kappa)}$ and $h \propto n^{-1/(2d-\bar\kappa)}$. Substituting into $C_\beta h^2$ gives (83); at $\bar\kappa = d-2$ the exponent is $2/(d+2)$, and the logarithm there enters $h^{q+2} \propto \Lambda \log n$ with $q+2 = (d+2)/2$. Monotonicity is that of $1/(2d-\bar\kappa)$, and $\kappa = 0$ makes F_0 free of h, so $h \to 0$ and the rate is $n^{-1/d}$. □

Table 1 tabulates the exponents. That $K = 1$ returns the empirical exponent, that the gain is bought with bias rather than information (Remark A.5), and why $2/(d+2)$ is a ceiling for a positive kernel and the known $s = 1$ minimax rate are discussed in Section 3. Two further points. For a μ_0 with no smoothness the bias is first order by Proposition D.1, the same optimisation gives $h_\star \asymp n^{-1/d}$, and the rate returns to $n^{-1/d}$: the entire advantage is the second-order cancellation of Proposition 4.1. In the other direction, a Hölder $D^2 f$ only shrinks the remainders of Proposition 4.1 and Lemma E.5 to $O(h_\star^{2+\gamma})$, and the rate does not move, since $\beta_1 = \Theta(h_\star^2)$ is a lower bound (Remark D.4).

The price of the gain is the one this paper charges everywhere. At the optimum the constant is $C_\beta^{(d-2)/(d+2)} C_F^{4/(d+2)}$, C_β as in (81), with $C_F \asymp c_d D (C_Q D)^{d/2-1}$, so the crossover at which (83) actually overtakes $c_d D n^{-1/d}$ sits at

$$n \;\gtrsim\; C_Q^{2d}, \tag{84}$$

the same exponential-in-d toll as (86) in Corollary G.5. Asymptotically better in every $d \geq 3$; usable where C_Q^{2d} is a sample size one has.

With the constants actually proved, that sample size is large. The proofs take $C_Q = \sqrt{d}$ (Definition B.2), the price of bounding a cell by a ball, so $C_Q^{2d} = d^d$, 10^{10} at $d = 10$, and on top of it the fluctuation carries the constant $C_{K,d}$ of Assumption 2, which pays for the worst-conditioned tuples

through $\Lambda_{K,d}$ of Proposition C.4. Measured, the threshold is nowhere near that. Since the condition is $k > C_Q^d$ with n cancelled, every gain observed at a small k bounds the effective constant, the value of C_Q at which the threshold would be exact: on the cosine torus the tuned generator beats the empirical measure at $n = 90$ with $k = 5$ at $d = 4$ (15%, 20 standard errors) and $k = 4$ at $d = 5$ (8%, 8 standard errors), so the effective C_Q is below 1.50 and 1.32 there, and the arity sweep at $d = 10$ with a 24-point pool puts it below 1.37. The effective crossover C_Q^{2d} is then at most 16 at $d = 5$ rather than $d^d = 3125$. The proof pays for cells as balls and for the worst-conditioned tuples; typical ones do far better. What the experiments support is the form of the threshold, exponential in d in the neighbourhood size with n cancelled, not its proven constant. An effective C_Q near 1.3 would put the default $k = 5$ below C_Q^d from $d \approx 6$, consistent with where Section 5.3 places the failure of SMOTE, though our runs bound C_Q only from above.

The constants themselves can be measured, and are.

Table 3: Measured constants. $C_{K,d}$ of Assumption 2 and b_K of Section A.4 Step 7 are for pairs on the uniform torus, at the k-NN scale $h_\star = (k/(nv_d))^{1/d}$ with $k = 10$, the n shown keeping $h_\star$ near 0.2; $C_{K,d}$ is the largest of the ratios over cell sides $s/h_\star \in \{\frac{1}{4}, \frac{1}{2}, 1, 2\}$ and b_K the largest cell ratio at $s = \frac{1}{4}$, against the proved $b_K \le K^2 = 4$. The bias coefficient c_K of $\bar{\nu} - \mu_0 = -c_K h_\star^2 \Delta f$ is measured for pairs ($K = 2$) on the cosine torus by a paired test function, extrapolated to $h_\star \to 0$, against the $(K-1)/((d+2)(K+1))$ of Proposition 4.1; $C_\beta = c_K \|\nabla f\|_{L^1}$. The crossover $n^\star = (C_Q^d)^{(d+2)/2}$, the condition $k > C_Q^d$ of Theorem A from $K = d - 1$ on, uses the effective $C_Q \in [1.3, 1.5]$ above.

d	n	$h_\star$	$C_{K,d}$	b_K	c_K measured	predicted	C_β meas./pred.	$n^\star$
3	600	.158	1.11	1.57	$.0745 \pm .0037$	.0667	.286/.256	7–21
4	2 000	.178	0.97	1.71	$.0567 \pm .0034$	.0556	.254/.249	23–130
5	6 000	.200	0.87	1.82	$.0513 \pm .0031$	.0476	.258/.240	99–1200
6	20 000	.214	0.81	1.81	$.0426 \pm .0036$	.0417	.235/.231	540–17000

Three readings. $C_{K,d}$ is of order one and decreases with d, its largest value coming from the finest cells, which is the regime Assumption 2 is about. b_K grows slowly and stays at half the proved K^2. The bias coefficient matches its closed form in every dimension, within 0.3 standard errors at $d = 4$ and $d = 6$ and within 12% at $d = 3$ and 5, which extends the circle check of Remark D.4 to $d \ge 3$. The crossover these constants imply, tens at $d = 4$ and hundreds to thousands at $d = 5, 6$, is consistent with the runs: a gain at $n = 90$ for $d = 4, 5$ and at $n = 203$ for $d = 3$. The proved $C_Q = \sqrt{d}$ would put it at 4096 for $d = 4$ and 2×10^9 for $d = 6$: it is the proved constant, not the mechanism, that is pessimistic.

Two columns were rerun independently. With a fixed radius, 2×10^6 draws per $h_\star \in \{0.1, 0.15, 0.2\}$ and a linear extrapolation in $h_\star^2$, c_K is 0.0676, 0.0553, 0.0475, 0.0417 at $K = 2$ and 0.0997, 0.0996, 0.0950, 0.0898 at $K = d$ for $d = 3, \dots, 6$, with standard errors 3–6 $\times 10^{-4}$, against the closed-form 0.0667, 0.0556, 0.0476, 0.0417 and 0.1000, 0.1000, 0.0952, 0.0893: within 1.4% at both arities, so the arity factor of Proposition 4.1 holds in $d \ge 3$ as well. The $C_{K,d}$ column reruns to 1.12, 0.99, 0.90, 0.83. At $K = 2, 3, 4, 6, 8, 12$ for $d = 4$ and 6 the same estimator, with 10^6 draws per $h_\star$, stays within 1.5% of the closed form at every arity; these are the markers of Figure 1B.

Remark G.4 (What the SMOTE bias is, and a comparison). For the k-NN generator, Lemma E.5 gives

$$\beta_1 \;\le\; \Big(\frac{K-1}{(d+2)(K+1)}\Big(1+\frac{2}{dK}\Big)\frac{\|f^{-2/d}\nabla f\|_{L^1}}{\omega_d^{2/d}} + C'_\partial\Big)\Big(\frac{k+1}{n_A}\Big)^{2/d} + o\Big(\Big(\frac{k}{n}\Big)^{2/d}\Big), \qquad (85)$$

squared in the k-NN radius rather than linear in it. Xia & Klusowski (2026) obtain a term of order $(k/N_1)^{1/d}$ in a total-variation bound on the excess risk of a classifier trained on SMOTE-augmented data. The analogous term here is the mixing bias β_1 of (1), the one component carrying no dependence on n, n_π or S; in W_1 and for a C^2 population distribution it is of the square of that order, because convex mixing with exchangeable weights cancels at first order and only the density gradient survives. The two quantities are not directly comparable, a total-variation bound on a risk against a W_1 distance; without smoothness of μ_0, Proposition D.1 gives the first order $(k/N_1)^{1/d}$ in W_1 as well.

Corollary G.5 (The price of a smoothing gain)**.** *For k-NN SMOTE the threshold* (30) *reduces to*

$$k \;>\; C_Q^d, \qquad (86)$$

in which n has cancelled: the ratio $\theta_{\mathrm{sm}}/\theta_{\mathrm{samp}}$ is a function of (k, d) alone and no amount of data creates it. If in addition the mixing bias is to satisfy $\beta_1 \le \eta D$, then $k \le \eta^{d/2} n$ by (85), so both can hold only when $n > (C_Q^2/\eta)^{d/2}$: a smoothing gain together with a small bias costs a sample size exponential in the dimension.

Proof. With $h_\star = D(k/n)^{1/d}$ (Lemma G.2) and $S \ge 1/n$, the threshold $h_\star > C_Q D S^{1/d}$ becomes $D(k/n)^{1/d} > C_Q D n^{-1/d}$, which is (86). Together with $k \le \eta^{d/2} n$ it gives $C_Q^d < \eta^{d/2} n$. □

H MANIFOLDS

Proof of Proposition 5.1. For $p, q \in \mathcal{M}$, $\mathrm{dist}(q - p, T_p\mathcal{M}) \le \|q - p\|^2/(2\tau)$ (Federer, 1959, Thm. 4.18). (i) Write $w = Y - Z_1 = \sum_{j \ge 2} u_j (Z_j - Z_1)$ and split it into its tangent part v and normal part at Z_1. The normal part has norm at most $\sum_j u_j \|Z_j - Z_1\|^2/(2\tau) \le h^2/(2\tau)$, and $\|v\| \le \|w\| \le h$. The unit-speed geodesic γ of $\mathcal{M}$ from Z_1 in the direction of v exists by compactness, and its acceleration is normal with norm at most $1/\tau$, the second fundamental form of $\mathcal{M}$ being bounded by $1/\tau$ (Niyogi et al., 2008, Prop. 6.1); so $\|\gamma(\|v\|) - Z_1 - v\| \le \|v\|^2/(2\tau)$, and $\mathrm{dist}(Y, \mathcal{M}) \le \|Y - \gamma(\|v\|)\| \le h^2/(2\tau) + h^2/(2\tau)$. (ii) Let $x \in \mathcal{M}$, $y = x + \sigma g$ and q a nearest point of $\mathcal{M}$ to y, so that $\|q - x\| \le 2\sigma\|g\|$. With P_N the projection on the normal space at x, $\mathrm{dist}(y, \mathcal{M}) = \|y - q\| \ge \sigma\|P_N g\| - \|P_N(q - x)\| \ge \sigma\|P_N g\| - 2\sigma^2\|g\|^2/\tau$. For $G \sim \mathcal{N}(0, I_d)$, $P_N G$ is standard Gaussian in $d - m$ dimensions, and Hölder's inequality gives $\mathbb{E}\|P_N G\| \ge (d-m)^{3/2}((d-m)(d-m+2))^{-1/2} \ge \sqrt{(d-m)/3}$, while $\mathbb{E}\|G\|^2 = d$; integrating against ρ, the mean distance is at least $(1/\sqrt{3} - 1/3)\,\sigma\sqrt{d-m} \ge \sigma\sqrt{d-m}/5$ once $2\sigma^2 d/\tau \le \sigma\sqrt{d-m}/3$. As μ_0 is carried by $\mathcal{M}$, every coupling moves each y to $\mathcal{M}$, so this is a lower bound on W_1. (iii) The proof of (4), with the mass bound $c_\mu r^m$, gives $W_1(\hat{\mu}_n, \mu_0) \ge c_e n^{-1/m}$, $c_e = \frac{1}{2}(2c_\mu)^{-1/m}$, and $W_1(\hat{\mu}_n, \hat{\mu}_n * \mathcal{N}_\sigma) \le \sigma\mathbb{E}\|G\| \le \sigma\sqrt{d}$. If $\sigma\sqrt{d} \le c_e n^{-1/m}/2$ the triangle inequality gives $W_1 \ge c_e n^{-1/m}/2$; otherwise (ii) gives $W_1 \ge \sigma\sqrt{d-m}/5 \ge \frac{c_e}{10}\sqrt{(d-m)/d}\, n^{-1/m}$. □

Remark H.1 (An intrinsic Theorem A, conjectured). By (i), every synthetic point lies within $h_\star^2/\tau$ of $\mathcal{M}$, so projecting $\widehat{\nu}_X$ onto $\mathcal{M}$ costs at most $h_\star^2/\tau$ in W_1, the order of the bias. We conjecture that Theorem A(i) then holds with $m \ge 3$ in place of d: for a density on $\mathcal{M}$ of class C^2 bounded below, $\kappa = \min(K - 1, m - 1)$ and $\bar{\kappa} = \min(\kappa, m - 2)$, tuned k-NN SMOTE attains $n^{-2/(2m-\bar{\kappa})}$. From $K = m - 1$ on this is the $s = 1$ rate on m-dimensional manifolds (Divol, 2022), which by (iii) no isotropic Euclidean kernel estimator reaches. It is not proved here.

Numerical checks of Proposition 5.1. On the unit sphere its bounds hold with room: over S^1, S^2, S^3, arities $2, 3, 5$ and $h_\star \in [0.02, 0.4]$ the worst mixing defect is 0.47–$0.49\,h_\star^2$, and the kernel's mean displacement is 0.88–$0.97\,\sigma\sqrt{d-m}$. The constants are explicit there. On a unit-curvature S^m a pair at geodesic distance r mixed with weight u sits $\frac{1}{2}u(1-u)r^2$ below the manifold, so the mean defect is $\frac{m}{12(m+2)}h_\star^2$, against $h_\star\sqrt{(d-m)/d}$ for an isotropic kernel of matching scale. Simulation on S^2 in $\mathbb{R}^d$ for $d = 4, 6, 10, 16, 24$ gives fitted exponents in $h_\star$ of 2.00–2.01 for mixing and 1.00 for the kernel, a mixing constant of 0.0413–0.0424 against $1/24$ at every d (after correcting for the $\approx 8\%$ of base points without a neighbour at $h_\star = 0.05$, which copy themselves and sit on the manifold; uncorrected the mixing exponent reads 2.04), and a kernel constant reaching 0.89 of $\sqrt{(d-2)/d}$ at $d = 4$ and 0.99 at $d = 24$. The defect ratio is 37 at $h_\star = 0.4$ and 490 at $h_\star = 0.05$, and the ordering survives in transport distance: at $d = 6$ with 4000-point clouds the kernel measures 0.315 and 0.170 at $h_\star = 0.4$ and 0.2, while mixing sits at the estimation floor, 0.067. The same holds on real data. On MNIST digits $0, 1, 3, 8$ in $\mathbb{R}^{784}$, measuring each synthetic point against a 12-dimensional tangent plane fitted to 40 held-out digits of its class, the kernel adds a normal displacement of $(0.99$–$1.02)h_\star$ at every locality $k = 2, \ldots, 80$, against $\sqrt{772/784} = 0.99$ predicted, while mixing adds none: its residual lies 16–24% below that of real digits, because averaging neighbours also averages their pixel noise, which makes the mixtures smoother, not more realistic. A Euclidean KDE thus pays a first-order off-manifold bias that mixing does not, and matching it requires the manifold-adapted kernels of Divol (2022). The structural differences off a manifold (convex constraints, labels, affine invariance, a scale-free k) are listed in Section 5.

I EXPERIMENTAL DETAILS

I.1 FIRST GROUP: ARITY SWEEPS AND THE TUNED RATE

Arity sweep: the torus population. Uniform distribution on the torus $[0,1)^d$ with the wrapped metric; $n = 400$ real points; neighbourhoods are the 24 nearest neighbours (a fixed radius gives empty neighbourhoods once $d \geq 6$ at any feasible n), from which $K-1$ partners are drawn without replacement; $N = 2 \times 10^4$ synthetic points per condition; $R = 10$ replicates. The divergence is the debiased Sinkhorn divergence with cost $\|x-y\|$, not its square, so that it approximates W_1 (the wrapped distance on the torus, the Euclidean one for the Gaussian; the same in every run of this appendix), at $\varepsilon = 10^{-3}$, whose residual entropic bias is $O(\varepsilon \log 1/\varepsilon)$ and is shared across K at fixed (N, ε). Within a replicate the sample and the reference are shared across all K, so the reported ratios are paired.

Table 4: Arity sweep on the torus ($\beta_1 \equiv 0$), raw and net of the estimation floor.

W_1 rel. to $K=2$	K=2	3	4	5	6	8	10	12
$d=4$	1	.924	**.917**	.925	.941	.965	.990	1.017
$d=6$	1	.919	**.901**	.901	.907	.924	.942	.961
$d=8$	1	.930	.910	**.907**	.910	.921	.936	.950
$d=10$	1	.939	.921	**.918**	.919	.928	.940	.951
net of the estimation floor, removed in quadrature								
$d=4$	1	.889	**.879**	.890	.914	.950	.986	1.024
$d=6$	1	.846	**.809**	.810	.822	.855	.892	.928
$d=8$	1	.833	.783	**.774**	.782	.811	.849	.884
$d=10$	1	.824	.766	**.755**	.761	.789	.826	.861

At every d but 6, where $K = 4$ and $K = 5$ tie, the winning arity wins in all ten replicates. Throughout this appendix $\pm$ is one standard deviation over the paired replicates, unless marked as a standard error, which is smaller by $\sqrt{R}$. Paired standard deviations on the top block are 0.006–0.013 at $d = 4$, 0.002–0.006 at $d = 6$ and 0.001–0.002 at $d = 8, 10$, over $R = 10$ replicates. The estimation floors, the same divergence between two independent uniform samples of size N, are 0.060, 0.148, 0.243, 0.337 for $d = 4, 6, 8, 10$, i.e. $55, 67, 75, 80\%$ of the $K = 2$ level; the lower block removes them in quadrature, $(\text{level}^2 - \text{floor}^2)^{1/2}$, as in Section 6 and the other tables. Since the floor does not depend on K it cannot move the minimum, and indeed the two blocks agree on $K^\star$ in every dimension. The overlap factor stays within 3% of 2.6×10^{-3} across all conditions.

Arity sweep: the Gaussian population. Identical in every respect except the population: $\mu_0 = \mathcal{N}(0, \sigma^2 I_d)$ with the Euclidean metric and $\sigma^2 = 1/24$, chosen so that $\mathbb{E}\|x-y\|^2 = d/12$, exactly its value on the torus with the wrapped metric; the same $n = 400$, 24-NN pool, $N = 2 \times 10^4$, $R = 10$ and $\varepsilon = 10^{-3}$, and the same pairing across K within a replicate. The reference is an i.i.d. sample from μ_0. The Gaussian lies outside the hypotheses of the rate theorems, which ask for bounded support, and of Proposition 4.1, which also asks for a density bounded below. What applies is Proposition A.8, with $\alpha_i = \sigma^2$, and the baseline through Remark A.12; the second-order bias expansion is used as a formal expansion with $C_\partial = 0$, since there is no boundary. The population is therefore an empirical test of the bias mechanism, and the $(K-1)/(K+1)$ gap it reproduces is a prediction checked beyond the proved scope, not an instance of a theorem.

Table 5: Arity sweep on the Gaussian ($\beta_1 > 0$), raw and net of the floor.

W_1 rel. to $K=2$	K=2	3	4	5	6	8	10	12
$d=4$	**1**	1.003	1.049	1.094	1.136	1.192	1.235	1.269
$d=6$	1	**.961**	.976	1.000	1.025	1.067	1.098	1.127
$d=8$	1	**.955**	.957	.972	.988	1.020	1.046	1.067
$d=10$	1	.953	**.948**	.957	.968	.992	1.014	1.032
net of the estimation floor, removed in quadrature								
$d=4$	**1**	1.004	1.069	1.131	1.188	1.264	1.321	1.367
$d=6$	1	**.924**	.953	1.000	1.048	1.125	1.181	1.232
$d=8$	1	**.885**	.891	.929	.970	1.047	1.108	1.158
$d=10$	1	.849	**.835**	.863	.901	.977	1.040	1.092

Paired standard deviations are 0.008–0.013 at $d = 4$, 0.009–0.014 at $d = 6$, 0.005–0.010 at $d = 8$ and 0.003–0.006 at $d = 10$; the estimation floors are $0.055, 0.131, 0.214, 0.297$, i.e. $54, 69, 77, 82\%$ of the $K = 2$ level. As on the torus the two blocks agree on $K^\star$ in every dimension. Two entries lie within one standard deviation of $K = 2$: $K = 3$ at $d = 4$ (1.003 ± 0.008) and $K = 5$ at $d = 6$ (1.000 ± 0.012). Every $K \geq 4$ at $d = 4$ is worse by at least 3.8 standard deviations, i.e. 12 standard errors.

The gap between the populations. Since $\beta_1 \equiv 0$ on the torus, the difference of the two tables at fixed (d, K) is the arity dependence of the mixing bias, up to whatever the populations do not share elsewhere. Normalised at $K = 12$ and compared with $\big[g(K) - g(2)\big]/\big[g(12) - g(2)\big]$ for $g(K) = (K-1)/(K+1)$, the arity factor of Proposition 4.1:

Table 6: The gap between the two populations against the arity factor of Proposition 4.1.

(Gaussian − torus)/gap at K=12	K=2	3	4	5	6	8	10	12
predicted, $(K-1)/(K+1)$	0	.33	.52	.65	.74	.87	.94	1
$d = 4$	0	.31	.52	.67	.77	.90	.97	1
$d = 6$	0	.25	.45	.60	.71	.86	.94	1
$d = 8$	0	.21	.40	.55	.67	.84	.94	1
$d = 10$	0	.17	.34	.49	.61	.80	.92	1

Nothing is fitted: the row and the prediction are computed independently. The gap at $K = 12$, in W_1 units on the common cost scale, is $0.0253, 0.0314, 0.0326, 0.0291$ for $d = 4, 6, 8, 10$, i.e. $\times 1.00, 1.24, 1.29, 1.15$, while $h_\star^2 \|\nabla f\|_{L^1}/(d+2)$ with $h_\star$ the mean 24-NN radius and $\|\nabla f\|_{L^1}$ evaluated globally gives $\times 1.00, 1.47, 1.71, 1.84$. The dependence on K is therefore quantitative evidence and the dependence on d only qualitative, which is what one should expect of a plug-in that replaces a supremum over the sample by an average.

Bootstrap baseline. $K = 1$ is a legal setting of the same generator ($Y = X_{\text{base}}$, so $\widehat{\nu}_X = \hat{\mu}_n$ up to multinomial fluctuation of the counts: the synthetic sample carries exactly n distinct values and $S = 2.69 \times 10^{-3}$ against $1/n = 2.5 \times 10^{-3}$). To keep an entropic artefact out of a baseline, this block uses exact optimal transport (Hungarian assignment) rather than Sinkhorn, at $n = 400$, $N = 2500$, $R = 4$, on the torus; it is the control for the Sinkhorn-based bootstrap ratios of the Gaussian grids below.

Table 7: Bootstrap baseline on the torus under exact optimal transport.

	floor	K=1	2	3	4	5	8
$d = 4$ level	.1015	.1466	.1285	.1251	.1270	.1270	.1299
rel. to K=1		1.000	.877	**.854**	.866	.867	.886
$d = 8$ level	.3159	.3834	.3534	.3448	.3421	.3422	.3454
rel. to K=1		1.000	.922	.899	**.892**	.892	.901

Net of the floor, removed in quadrature, the best arity is 31% ($d = 4$) and 40% ($d = 8$) below pure resampling. A smoothed bootstrap was run alongside and is not reported: on a group with an invariant uniform distribution an isotropic kernel is unbiased by construction, so this population cannot compare mixing with kernel smoothing either way. Any advantage of convex mixing over kernel smoothing would have to come from the directions it blurs in and from combinations staying inside the hull of the sample, neither of which a flat uniform population can exhibit, and both of which are what the high-dimensional SMOTE literature reports missing (Blagus & Lusa, 2013).

Robustness: the entropic parameter and the reference. Both checks use $N = 16384$ with $K \in \{2, 3, 4, 5, 6, 8\}$ and a Sobol reference, $R = 3$ replicates for the ε-sweep and $R = 2$ for the reference comparison.

Table 8: Robustness of the arity sweep to the entropic parameter.

W_1 rel. to $K=2$		K=2	3	4	5	6	8
$\varepsilon = 3 \times 10^{-4}$	$d = 4$	1	.928	**.919**	.925	.944	.972
	$d = 10$	1	.942	.926	**.922**	.924	.932
$\varepsilon = 10^{-3}$	$d = 4$	1	.930	**.921**	.927	.945	.974
	$d = 10$	1	.943	.927	**.923**	.925	.933
$\varepsilon = 3 \times 10^{-3}$	$d = 4$	1	.934	**.926**	.933	.951	.980
	$d = 10$	1	.945	.929	**.925**	.927	.935

A tenfold change of ε moves $K^\star$ nowhere and the level at the minimum by $+0.77\%$ ($d = 4$) and $+0.30\%$ ($d = 10$), monotonically and in the direction over-smoothing predicts. Replacing the i.i.d. reference by a scrambled Sobol sequence of the same size lowers the estimation floor from 0.0632 to 0.0599 at $d = 4$ (-5.2%) but only from 0.34390 to 0.34357 at $d = 10$ (-0.1%), since a low-discrepancy sequence needs $N \gg 2^d$ to be in its regime; the $K = 4$ ratio is $.926 \pm .008$ (i.i.d.) against $.920 \pm .007$ (Sobol) at $d = 4$ and $.928 \pm .001$ against $.927 \pm .001$ at $d = 10$, over two replicates only. We record it as a lever that was tried and rejected rather than as a design choice.

Two diagnostics accompany every condition. The overlap factor stays in $[2.67, 2.91] \times 10^{-3}$ across all K, both populations and $K = 1$ included: the design fixes $w_i \approx 1/n$ whatever the arity, so nothing in these curves is reuse. And the standard deviation of the synthetic sample from $K = 2$ to $K = 6$ moves from 0.2914 to 0.2916 on the torus but from 0.1876 to 0.1720 on the Gaussian: the shrinkage of a K-fold mix towards local centroids, invisible where translation invariance pins the marginal and plain to see where it does not.

Concentration of the Dirichlet push-forward. For K points i.i.d. uniform in a unit ball, let σ_K be the mean distance from $\sum_j U_j Z_j$ to the centroid of the Z_j, normalised at $K = 2$ (4×10^5 draws per cell).

Table 9: Concentration of the Dirichlet push-forward: σ_K is dimension-free, and the penalty $\sigma_K^{-(d/2-1)}$ it induces in (31) crosses one between $K = 4$ and $K = 5$.

σ_K	K=2	3	4	5	6	8	10	12	16
$d = 2$	1.000	1.068	1.032	.981	.934	.850	.782	.728	.645
$d = 4$	1.000	1.064	1.029	.979	.930	.846	.780	.727	.644
$d = 8$	1.000	1.062	1.027	.978	.931	.848	.781	.728	.647
$d = 12$	1.000	1.063	1.028	.979	.931	.849	.783	.729	.648
penalty $\sigma_K^{-(d/2-1)}$ in (31)									
$d = 4$	1	0.94	0.97	1.02	1.08	1.18	1.28	1.38	1.55
$d = 6$	1	0.88	0.94	1.04	1.15	1.39	1.64	1.89	2.39
$d = 8$	1	0.84	0.92	1.07	1.24	1.64	2.10	2.59	3.70
$d = 10$	1	0.79	0.90	1.09	1.33	1.94	2.69	3.55	5.71

The rows of the top block agree to three decimals: the relative spread of the Dirichlet push-forward is dimension-free, peaking at $K = 3$ and falling by 27% at $K = 12$. The penalty it induces in (31) is not dimension-free, but its crossing of one still sits between $K = 4$ and $K = 5$ in every dimension, which is where Figure 1A puts $K^\star$ on the torus.

The locality lever: prediction and outcome. The two mechanisms that shape the arity sweeps, the concentration of the weights and the bias, scale differently in the locality: σ_K is a property of the Dirichlet weights alone, while $\beta_1 = \Theta(h_\star^2)$. Contracting the neighbourhood pool from 24 to 6 therefore divides $h_\star^2$ by $4^{2/d}$, a factor two at $d = 4$ but only 1.32 at $d = 10$, and must move $K^\star$ to the right at $d = 4$ while leaving $d = 10$ nearly fixed. If $K^\star$ did not move at $d = 4$, or moved left, reading the gap between the populations as β_1 would be wrong. Both grids use $N = 16384$, $R = 6$, a Sobol reference and the same seeds, so the $K = 1$ column is identical across them and they differ in exactly one parameter.

Table 10: The locality lever: the two neighbourhood pools.

W_1 rel. to $K=2$		K=1	2	3	4	5	6
pool = 24	$d=4$	1.273	**1**	1.004	1.052	1.093	1.137
	$d=6$	1.234	1	**.962**	.977	1.000	1.027
	$d=10$	1.178	1	.955	**.951**	.958	.970
pool = 6	$d=4$	1.138	1	**.979**	.997	1.024	1.057
	$d=6$	1.134	1	**.962**	.968	.984	1.007
	$d=10$	1.121	1	.959	**.954**	.960	.972

$K^\star$ moves $2 \to 3$ at $d=4$, stays at 3 at $d=6$ and at 4 at $d=10$, as predicted. In level terms the contraction lowers the $K=6$ ratio by 0.080, 0.020 and -0.002 for $d=4,6,10$ (standard errors 0.009, 0.008, 0.002), against 0.097, 0.044 and 0.012 obtained by scaling the measured gap between the populations at $K=6$ by $1-4^{-2/d}$; and the displacement, normalised at $K=6$, reads $.32, .68, .87, 1$ at $d=4$ against $.44, .70, .88, 1$ from $(K-1)/(K+1)$, agreeing from $K=4$ on and undershooting at $K=3$, where the displacement is a few thousandths. The pool-24 rows also replicate the main Gaussian table above to within half a standard deviation, at a different N, R and reference. Against the $K=1$ bootstrap, the best arity sits at $0.786, 0.780, 0.807$ of the resampling level at pool 24 and at $0.860, 0.848, 0.850$ at pool 6 ($0.73, 0.65, 0.56$ and $0.83, 0.77, 0.67$ net of the floor): the contraction that buys a smaller bias sells resolution, which is the two-sided role of ε in Corollary G.5 within a single grid. One caveat on the small grid: at pool 6 and $K=6$ the five partners are drawn from six candidates, so that column carries almost no selection variability.

Expanding the pool. If the torus optimum stays below d because the ratio $G = \theta_{\mathrm{sm}}/\theta_{\mathrm{samp}} \asymp k^{1/d}/C_Q$ of (31) is close to one, so that the concentration penalty outweighs it, widening the pool should favour higher arity, and $W_1(K{=}6)/W_1(K{=}4)$ at $d=6$ should decrease over $k = 24, 96, 384$ at fixed locality ($n = 20k$). It fell from 1.0054 to 1.0011 (standard errors 0.0024 and 0.0010), in the predicted direction but by 1.7 standard errors only and not monotonically (1.0002 at $k=96$): at fixed locality n grows with k and pairs approach the floor, which compresses every arity difference. The test is underpowered, and this explanation is consistent with it, not demonstrated by it.

The tuned rate. The tuned generator ($K=3$, $k=\lceil n^{0.4}\rceil$, the rule (2) of Theorem A at $d=3$, without the split) is compared with the empirical measure over $n = 203, \dots, 3250$, with 65536-point clouds, $R=8$ replicates, one reference cloud and one nested pool of real points per replicate, and the floor removed by the quadrature rule measured $\approx (\text{true}^2 + \text{floor}^2)^{1/2}$, which holds to about 1% on pure translations, where W_1 is the shift. On the cosine torus $f = \prod_j (1 + \frac{1}{2}\cos 2\pi x_j)$, which meets the hypotheses of Theorem A on the density, with $\beta_1 > 0$ and no boundary, the synthetic level falls as $n^{-0.388}$ against $n^{-0.400}$ predicted, and on the flat torus, where $\beta_1 \equiv 0$, as $n^{-0.391}$. The ratio to the empirical measure falls as $n^{-0.032}$ and $n^{-0.037}$ (standard errors 0.004 and 0.005), eight standard errors from zero but half the predicted $n^{-0.067}$. Under the quadrature rule the shortfall sits on the empirical side, whose corrected slope is -0.355 in all three populations, the Gaussian included, against the $n^{-1/3}$ it provably has; Table 11 below shows how far that reading depends on the distribution. On the Gaussian, outside Theorem A since its support is unbounded and its density not bounded below, the ratio is flat, $+0.002$ (standard error 0.008): tuned mixing wins a constant 17–19% at every n but no rate. The tuned k beats the untuned $k=5$ at every n in all three populations.

The gap $2/(d+2) - 1/d$ is largest near $d=5$, so the cosine torus was rerun at $d=4$ and 5 with $K=d$, still super-critical, over $n = 90, \dots, 1448$ ($N = 65536$ and 131072). The synthetic level falls as $n^{-0.310}$ and $n^{-0.280}$, against $n^{-0.333}$ and $n^{-0.286}$, and the ratio as $n^{-0.034}$ and $n^{-0.049}$ (standard errors 0.003), 12 and 15 standard errors from zero, against $n^{-0.083}$ and $n^{-0.086}$. Under the quadrature rule the empirical side overshoots its proven slope by 0.023, 0.024 and 0.026 at $d=3,4,5$. Every synthetic cloud is above the $N \gtrsim n^{2d/(2d-\bar\kappa)}$ of Corollary 3.1 at the largest n: that threshold is 1.6×10^4, 1.6×10^4 and 3.3×10^4 at $d=3,4,5$, against 65536, 65536 and 131072 points.

How much the floor decides. Every level above is measured against a reference cloud, and the floor, the same divergence between two independent N-point samples of μ_0, is 18–45% of the empirical level, growing with n. The quadrature rule is exact for a translation, not in general, so Table 11 refits the three slopes with the floor removed as $(\text{level}^p - \text{floor}^p)^{1/p}$: not at all ($p=\infty$), in quadrature

($p = 2$), linearly ($p = 1$), and with the p at which the empirical side has exactly its asymptotic slope $-1/d$ ($p = 3.0, 3.5, 4.5$). The ratio slopes are weighted fits as above, with standard errors 0.002–0.004.

Table 11: Slopes in $\log n$ of the synthetic level, the empirical level and their ratio on the cosine torus, by how the estimation floor is removed.

	$d = 3$			$d = 4$			$d = 5$		
floor removed	syn	emp	ratio	syn	emp	ratio	syn	emp	ratio
predicted	$-.400$	$-.333$	$-.067$	$-.333$	$-.250$	$-.083$	$-.286$	$-.200$	$-.086$
not at all	$-.333$	$-.324$	$-.008$	$-.260$	$-.245$	$-.012$	$-.233$	$-.198$	$-.027$
calibrated on emp	$-.353$	$-.333$	$-.019$	$-.272$	$-.250$	$-.020$	$-.238$	$-.200$	$-.032$
in quadrature	$-.388$	$-.356$	$-.032$	$-.310$	$-.274$	$-.034$	$-.280$	$-.226$	$-.049$
linearly	$-.529$	$-.461$	$-.069$	$-.431$	$-.363$	$-.068$	$-.393$	$-.305$	$-.084$

Three readings. The sign is robust: the ratio falls under every correction and in every d, by 2.7 standard errors at least (no correction, $d = 3$). The size is not: at $d = 3$ the ratio slope runs from -0.008 to -0.069, the prediction included, because the synthetic level, lower than the empirical one, carries the larger share of floor, and the predicted slope is a small difference of two exponents, $2/(2d - \bar{\kappa})$ and $1/d$, which a few percent on either side halves. The weighted fit is not the issue: an unweighted one moves the quadrature ratios by at most 0.005. And the correction that gives the benchmark its known exponent puts the synthetic side at 82–88% of its predicted exponent and the ratio at a quarter to a third of its prediction, so part of the shortfall is the generator's own pre-asymptotic regime: k runs over 4–26 and the locality $h_\star$ changes by a factor 1.5–1.8 only across each grid. Settling the size needs a floor well below the signal, about 10^7 points at $d = 3$, beyond these runs.

I.2 Second group: kernels, manifolds and arity

This subsection documents the second group of experiments of Section 6. Table 13 lists the runs and Table 12 summarises them at the largest n.

Table 12: W_1 of each generator over that of the sample at the largest n (Table 13), floor removed in quadrature: the numbers of Figure 1D. m is the intrinsic dimension. †At the estimation floor.

	B: bumps in $[0,1]^d$			C: S^3 bent into $\mathbb{R}^d$			D: T^4 in $\mathbb{R}^{16}$	
	$d = 3$	4	5	$d = 8$	16	32	unif.	cos
SMOTE, $k = 5$	0.92	0.88	0.86	0.93	0.93	0.93	0.91	0.90
rule k of (2), $K = m$	0.83	0.85	0.86	0.79	0.78	0.78	0.83	0.84
tuned mixing, $K = m$	0.83	0.84	0.82	0.72	0.71	0.69	0.77	0.79
tuned mixing, $K = 2$	0.79	0.76	0.68	0.67	0.67	0.65	0.73	0.73
Gaussian kernel	0.56	0.52	0.54	0.93	0.97	0.98	0.93	0.93
Gaussian, reflected at the faces	0.54	0.48	0.50	–	–	–	–	–
Gaussian on the PCA subspace	–	–	–	0.76	0.76	0.76	0.84	0.84
Gaussian in the true chart	–	–	–	0.27	0.23	0.29	0.00†	0.32

Populations. The three families of Section 6 are arenas B, C and D; arena A is used only to motivate B. *Arena A* is the cosine torus of Figure 1C, $f \propto \prod_j (1 + a \cos 2\pi x_j)$ with $a = 0.5$ and the wrapped metric. *Arena B* lives in $[0,1]^d$. With probability 0.3 a point is uniform; otherwise it comes from one of four Gaussian bumps of equal weight and standard deviation 0.12, centred at points drawn once in $[0.2, 0.8]^d$ and truncated to the cube. *Arena C* carries the density $\propto 1 + ax_1$, $a = 0.5$, on $S^3 \subset \mathbb{R}^4$ and maps it into $\mathbb{R}^d$ by $F(x) = Q_1 x + Q_2 g(x)$, where $[Q_1\, Q_2]$ is a random orthogonal matrix and $g_j(x) = \beta(\omega\sqrt{J})^{-1} \sin(\langle \omega_j, x\rangle + \phi_j)$ for $j \le J = d - 4$, with $|\omega_j| = \omega = 2$, random directions and phases, and bend $\beta \in \{0, 0.3, 1\}$. Since $|F(x) - F(y)|^2 = |x - y|^2 + |g(x) - g(y)|^2$, the map F expands distances, and its image spans $\mathbb{R}^d$. *Arena D* draws θ on the flat torus $[0,1)^4$, uniformly or with density $\prod_j (1 + a \cos 2\pi\theta_j)$, maps it to the Clifford torus $u(\theta) = (2\pi)^{-1}(\cos 2\pi\theta_j, \sin 2\pi\theta_j)_{j \le 4} \in \mathbb{R}^8$, whose intrinsic metric is the flat one, and then into $\mathbb{R}^{16}$ by the F of arena C with $\beta = 0.3$, the amplitude scaled by the circle radius $(2\pi)^{-1}$. In the uniform arm the only mixing bias is the second-order defect of chords.

Generators. The *sample* is the empirical measure of the n points. *Mixing* with arity K and locality k draws a base point uniformly among the n, $K-1$ distinct partners uniformly among its k nearest neighbours, and weights $U \sim \mathrm{Dir}(1,\dots,1)$; the split of Theorem A is not used. SMOTE is $K=2$, $k=5$, and the rule is $k=\lceil n^{(m-\bar\kappa)/(2m-\bar\kappa)}\rceil$ of (2) with the intrinsic m in place of d. The *Gaussian kernel* adds σG to the base point, with G standard normal in $\mathbb{R}^d$. The *reflected kernel* (B) folds the result back into the cube at the faces. The *PCA kernel* (C, D) adds $\sigma P_V G$, where V spans the top q principal directions of the n points, $q=m+1$ in C and $2m$ in D, the dimension of the linear span of the unbent manifold. The *oracle chart kernel* (C, D) takes a Gaussian step of scale σ along the manifold (the exponential map of S^3, or a wrapped step on the flat torus) and then applies F. Figure 1 calls the last three "Gaussian, reflected at faces", "Gaussian on PCA subspace" and "Gaussian in true chart".

Measurement. W_1 is the debiased Sinkhorn divergence with cost $|x-y|$ and $\varepsilon=10^{-3}$, annealed from $10^3\varepsilon$ at decay 0.98, on the dense cost matrix. Against a converged reference, annealing moves the ratios by at most 0.2% in every arena. A generator is scored against a reference cloud of N points of μ_0, and so is a second fresh cloud of μ_0, whose score f is the estimation floor of that replicate. A score s is corrected to $t=(s^p-f^p)^{1/p}$ with $p=2$ (quadrature, primary), $p=\infty$ (no correction), or the calibrated p at which the corrected level of the sample falls as $n^{-1/m}$. In B no such p exists, because even the uncorrected level of the sample falls faster than $n^{-1/d}$ there. All generators in a (replicate, n) cell share base indices, Dirichlet and Gaussian draws. A ratio is the mean over the R replicates of the per-replicate ratio of corrected scores, with standard error $\mathrm{sd}/\sqrt{R}$. A slope is the weighted least-squares slope of $\log$ ratio on $\log n$, with weights $(\text{ratio}/\text{se})^2$.

Tuning. Scales are tuned on 4 replicates (2 in D) at $N=16384$, disjoint from the evaluation replicates. For each n and generator, the scale minimising the mean log ratio to the sample is selected from a grid: $k\in\{\frac14,\frac12,1,2,4\}$ times the rule for mixing, and $\sigma\in\{\frac{1}{16},\dots,4\}$ times a centre for a kernel, the centre being the Gaussian whose intrinsic spread matches that of a Dirichlet pair at the rule's k-NN radius. A minimum on an edge extends the grid by factors of two, up to three times per side; $k=n-1$ is a hard edge. Arena D was warm-started, each grid being the multiple selected in C at $d=16$ times $\{\frac12,1,2\}$, still extended at its edges (Table 23). No selection is left on an extendable edge; $k=n-1$ is selected in 3 of 92 mixing cells, all at $n\le 181$. At full N the selected scale is checked against its halves and doubles in B at $d=4$ and in C at $d=16$ (Table 24).

Checks. Tables 21 and 22 test the predictions of Sections 3 to 5 on these populations. On S^3: the ratio of $K=3$ to the sample falls faster with n than that of the Gaussian kernel (C1) and than that of the PCA kernel (C4); $K=3$ is below the sample at every n with the floor left in, and its ratio falls with n (C2); $K=3$ beats the Gaussian kernel at every n (C3); and, across d, the kernel's ratio to the sample rises (C1b) while that of $K=3$ to the kernel falls (C3b). On T^4: $K=4$ beats the Gaussian kernel (D1a) and the PCA kernel (D1b) at every n; $K=4$ beats pairs at every n, as its exponent favours it (D2; D2′ for $K=3$); the lead of pairs over $K=4$ is larger with the cosine density, whose bias is larger (D3); and $K=4$ is below the sample at every n with the floor left in (D4). On the bumps, the kernel against mixing (B-level) and the slope of each ratio to the sample against its exponent (B-rate). A check "at every n" passes if it holds at every n by two standard errors, fails if the opposite holds by two standard errors at some n, and is otherwise inconclusive. A check on a slope or a difference passes or fails when it clears zero by two standard errors.

Why a bounded domain. The cosine torus (arena A) is a poor test of mixing against a kernel: there a tuned kernel beats every generator, and tuned mixing drifts to a global k (13 of 18 selections at $k\ge 0.85n$, 10 of them at $k=n-1$), since with wrapped offsets global mixing is itself a wide smoother. For the same reason arena A says little about arity; there $K=m$ beats pairs at the smallest n (pairs over $K=m$: 1.08–1.48). Arena B replaces it with a bounded domain. Pairs against the PCA kernel and the chart kernel, and $K=3$ against $K=4$, are descriptive comparisons, with no check attached.

Results. Tables 14 to 17 give every generator at every n, Tables 18 and 19 the paired head-to-head ratios, Table 20 the slopes against the exponents, Tables 21 and 22 every check under the three corrections, and Figure 2 all of it at a glance. Four observations complement Section 6.

- *B.* The plain Gaussian kernel is within a few hundredths of the reflected one (0.52–0.56 against 0.48–0.54 at the largest n), so its leak across the faces costs little here. The ratio of the reflected kernel to the sample falls with n at $d = 3, 4$ as its exponent predicts once the floor is removed (without the correction this check fails; B-rate in Table 22). In B the tuned k moves towards the rule as n grows, from 2–11 times it at the smallest n to 1–8 times at the largest (Table 23), so tuning helps most at small n, and the ratios of tuned mixing to the sample do not fall with n; that of the rule falls in every d, by at most 28% of its predicted slope (Table 20).
- *C.* The bend barely matters: at $\beta = 1$ the levels move by a few hundredths and every check has the verdict it has at $\beta = 0.3$ and the same d. The selected k of pairs and of $K = 3$ are the same multiples of the rule at every d and bend except in one cell (Table 23), so the locality of mixing is set by the intrinsic geometry, while the bandwidth of the kernel shrinks with the codimension. The slope checks C1 and C2 need the floor removed; C2 is inconclusive at $d = 16$ and on the unbent sphere, and C1 on the unbent sphere.
- *D.* In the uniform arm the oracle is uninformative: its tuned bandwidth covers the torus, and it regenerates the population. In the cosine arm it is about as far ahead as in C (0.32). $K = 3$, the smallest arity with $\bar{\kappa} = 2$, is the best mixing generator at the smallest n in both arms and close to pairs after. The margin of pairs over $K = 4$ is smaller in the uniform arm (D3; $+0.011 \pm 0.003$ in mean log ratio), the direction a smaller bias scale predicts.
- *PCA kernel.* Over the seven manifold runs pairs are level with or ahead of it at the smallest n and ahead after (0.85–1.00).

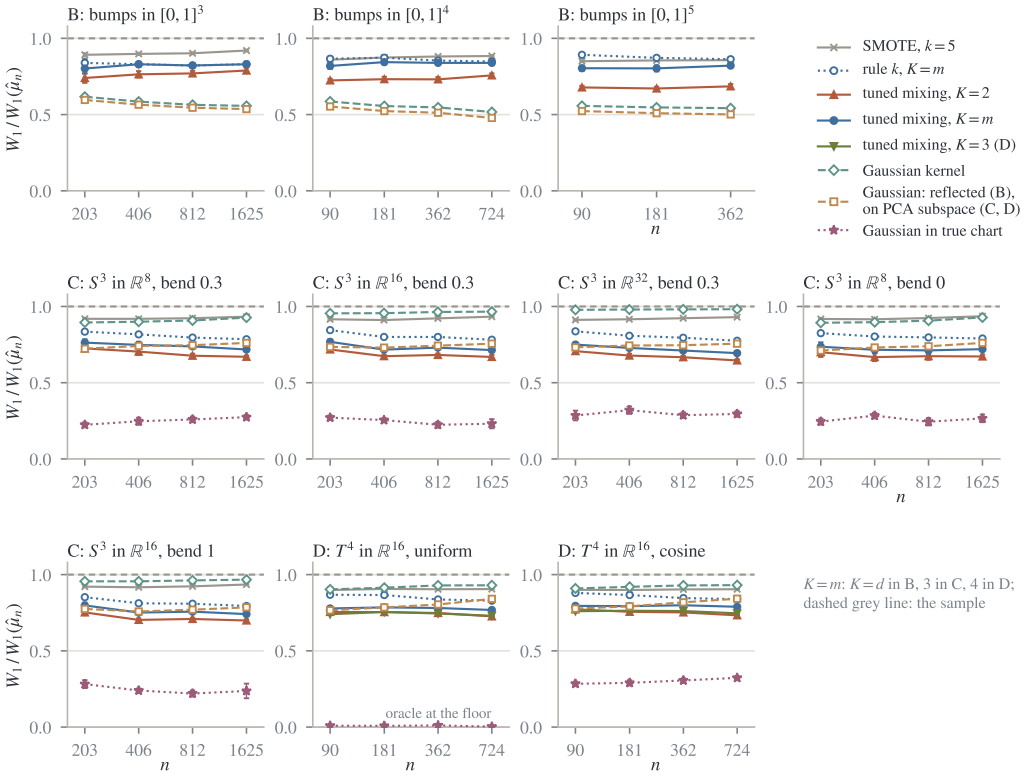

Figure 2: Every population of the second group: W_1 of each generator over that of the sample against n, floor removed in quadrature, one standard error.

Arity at equal $\bar{\kappa}$. Write the error of tuned mixing at locality h as a bias $b_K h^2$ plus a fluctuation. At $\bar{\kappa} = m - 2$ every level of the chaining of Section 3 below h contributes the same amount, and the fluctuation is $A_K n^{-1/2} h^{-(m-2)/2}$ up to a logarithm. Then

$$\min_h \left\{ b_K h^2 + A_K n^{-1/2} h^{-(m-2)/2} \right\} \asymp b_K^{(m-2)/(m+2)} A_K^{4/(m+2)} n^{-2/(m+2)},$$

the tuned rate of Theorem A at $\bar{\kappa} = m - 2$. Two arities with the same $\bar{\kappa}$ therefore differ by the factor $(b_K/b_{K'})^{(m-2)/(m+2)}$ if their fluctuation constants agree. The generator is the k-NN rule, for which $b_K \propto \frac{K-1}{K+1}(1 + \frac{2}{mK})$ by Lemma E.5 with m in place of d. At $m = 3$ every $K \geq 2$ has $\bar{\kappa} = 1 = m - 2$, and pairs against $K = 3$ should give $(8/11)^{1/5} = 0.938$. At $m = 4$, $K = 3$ and $K = 4$ have $\bar{\kappa} = 2 = m - 2$, and $K = 3$ against $K = 4$ should give $(70/81)^{1/3} = 0.952$. The fixed-radius factor $(K - 1)/(K + 1)$ of Proposition 4.1 would give $(2/3)^{1/5} = 0.922$ and $(5/6)^{1/3} = 0.941$. The measured ratios are 0.92–0.95 over every n of the six runs with $m = 3$ (B at $d = 3$ and the five C runs), with mean 0.937 over 24 cells, and 0.95–0.96 over every n of both D arms, with mean 0.954 over 8: the k-NN values, not the fixed-radius ones. In the uniform arm of D the density has no gradient on the torus, the bias is the off-manifold term of Proposition 5.1(i), and the k-NN factor has no reason to apply; both arms nevertheless give 0.95. The check is post hoc and heuristic: Theorem A is proved for full-dimensional densities, and the logarithm at $\bar{\kappa} = m - 2$ is ignored. When $\bar{\kappa}$ differs, as for pairs against $K = 4$ at $m = 4$, the same balance gives an error ratio of $K = 4$ over pairs proportional to $\lambda^{4/21} n^{-1/21}$ in the bias scale λ. This is the direction D3 tests. The exponent does not show at $n \leq 724$; the trend even runs the other way, the ratio of pairs to $K = 4$ having slope -0.012 ± 0.001 in n in the uniform arm rather than $+1/21$. At $n = 90$, on the other hand, $K = 3$ beats pairs by about 2% in both arms.

Robustness. *Floor removal.* Every level check (C3, D1, D2, D4 and B-level) has the same verdict without any correction (Tables 21 and 22): raw ratios include the floor in numerator and denominator, which pulls them towards one and can only shrink a gap. The slope checks need the floor removed; without it C1 passes only at $d = 8$ and C2 nowhere, which is why we claim the sign and level of the gain, not its exponent. *Tuning.* The neighbour check finds no halved or doubled scale better than the selected one by more than 1.7% in B at $d = 4$ and 0.1% in C at $d = 16$ (Table 24). *Replicates.* The unbent sphere and the bend-1 run of C used $R = 4$; every other run used $R = 8$.

Table 13: The runs of the second group. m: intrinsic dimension; R: evaluation replicates (tuning used disjoint ones); floor: the estimation floor over the sample's W_1 at the largest n.

run	population	m	d	n	N	R	floor
B, $d = 3$	B: bumps in $[0,1]^3$	3	3	203–1625	32768	8	43%
B, $d = 4$	B: bumps in $[0,1]^4$	4	4	90–724	32768	8	43%
B, $d = 5$	B: bumps in $[0,1]^5$	5	5	90–362	32768	8	43%
C, $d = 8$	C: S^3 in $\mathbb{R}^8$, bend 0.3	3	8	203–1625	32768	8	44%
C, $d = 16$	C: S^3 in $\mathbb{R}^{16}$, bend 0.3	3	16	203–1625	32768	8	44%
C, $d = 32$	C: S^3 in $\mathbb{R}^{32}$, bend 0.3	3	32	203–1625	32768	8	44%
C, $d = 8$, unbent	C: S^3 in $\mathbb{R}^8$, bend 0	3	8	203–1625	32768	4	43%
C, $d = 16$, bend 1	C: S^3 in $\mathbb{R}^{16}$, bend 1	3	16	203–1625	32768	4	44%
D, uniform	D: T^4 in $\mathbb{R}^{16}$, uniform	4	16	90–724	32768	8	45%
D, cosine	D: T^4 in $\mathbb{R}^{16}$, cos	4	16	90–724	32768	8	45%

Table 14: W_1 of each generator over that of the sample, floor removed in quadrature, one standard error over paired replicates; raw: the same ratio at the largest n with the floor left in. Arena B.

B: bumps in $[0,1]^3$	$n = 203$	$n = 406$	$n = 812$	$n = 1625$	raw
SMOTE, $k = 5$	0.892±.011	0.898±.007	0.902±.006	0.920±.006	0.935
mixing, $K = 2$	0.740±.032	0.764±.021	0.770±.016	0.789±.018	0.832
mixing, $K = 3$	0.802±.034	0.830±.015	0.822±.013	0.830±.016	0.864
rule k, $K = 3$	0.839±.025	0.830±.015	0.822±.013	0.830±.016	0.864
Gaussian kernel	0.617±.015	0.585±.017	0.564±.013	0.557±.012	0.661
reflected kernel	0.596±.014	0.565±.017	0.545±.013	0.536±.013	0.647
B: bumps in $[0,1]^4$	$n = 90$	$n = 181$	$n = 362$	$n = 724$	raw
SMOTE, $k = 5$	0.860±.008	0.874±.007	0.881±.005	0.884±.006	0.906
mixing, $K = 2$	0.724±.011	0.732±.016	0.731±.015	0.757±.013	0.808
mixing, $K = 4$	0.819±.020	0.844±.016	0.839±.014	0.839±.013	0.871
rule k, $K = 4$	0.866±.014	0.873±.013	0.854±.011	0.848±.011	0.878
Gaussian kernel	0.586±.013	0.556±.011	0.547±.010	0.517±.008	0.635
reflected kernel	0.553±.012	0.523±.011	0.512±.011	0.478±.009	0.610
B: bumps in $[0,1]^5$	$n = 90$	$n = 181$	$n = 362$		raw
SMOTE, $k = 5$	0.850±.006	0.855±.006	0.857±.005		0.886
mixing, $K = 2$	0.678±.011	0.671±.011	0.685±.015		0.755
mixing, $K = 5$	0.804±.017	0.803±.020	0.821±.011		0.858
rule k, $K = 5$	0.892±.010	0.872±.012	0.863±.009		0.890
Gaussian kernel	0.557±.012	0.547±.010	0.542±.007		0.654
reflected kernel	0.523±.011	0.509±.009	0.501±.006		0.627

Table 15: W_1 of each generator over that of the sample, floor removed in quadrature, one standard error over paired replicates; raw: the same ratio at the largest n with the floor left in. Arena C, bend 0.3.

C: S^3 **in** $\mathbb{R}^8$**, bend 0.3**	$n = 203$	$n = 406$	$n = 812$	$n = 1625$	raw
SMOTE, $k = 5$	0.919±.006	0.919±.004	0.923±.004	0.933±.003	0.946
mixing, $K = 2$	0.725±.021	0.704±.022	0.677±.013	0.670±.008	0.745
mixing, $K = 3$	0.763±.022	0.748±.021	0.738±.012	0.718±.008	0.780
rule k, $K = 3$	0.835±.012	0.817±.010	0.796±.008	0.785±.005	0.831
Gaussian kernel	0.895±.002	0.900±.002	0.908±.001	0.927±.001	0.942
PCA kernel	0.723±.007	0.741±.006	0.746±.004	0.761±.003	0.812
oracle chart kernel	0.224±.016	0.247±.023	0.259±.018	0.274±.014	0.502
C: S^3 **in** $\mathbb{R}^{16}$**, bend 0.3**	$n = 203$	$n = 406$	$n = 812$	$n = 1625$	raw
SMOTE, $k = 5$	0.916±.006	0.911±.002	0.922±.002	0.933±.002	0.946
mixing, $K = 2$	0.718±.016	0.674±.010	0.682±.009	0.669±.013	0.745
mixing, $K = 3$	0.768±.016	0.717±.010	0.730±.009	0.714±.012	0.778
rule k, $K = 3$	0.845±.011	0.800±.006	0.800±.005	0.782±.006	0.829
Gaussian kernel	0.955±.001	0.956±.000	0.964±.000	0.966±.000	0.973
PCA kernel	0.735±.006	0.730±.003	0.743±.002	0.756±.003	0.809
oracle chart kernel	0.271±.014	0.255±.011	0.224±.019	0.231±.030	0.490
C: S^3 **in** $\mathbb{R}^{32}$**, bend 0.3**	$n = 203$	$n = 406$	$n = 812$	$n = 1625$	raw
SMOTE, $k = 5$	0.912±.006	0.916±.004	0.923±.002	0.930±.002	0.944
mixing, $K = 2$	0.707±.012	0.678±.012	0.667±.009	0.646±.008	0.728
mixing, $K = 3$	0.750±.014	0.728±.013	0.711±.009	0.694±.008	0.763
rule k, $K = 3$	0.837±.011	0.808±.007	0.795±.006	0.775±.006	0.824
Gaussian kernel	0.978±.001	0.980±.000	0.981±.000	0.982±.000	0.985
PCA kernel	0.732±.007	0.744±.004	0.746±.005	0.756±.003	0.809
oracle chart kernel	0.285±.032	0.320±.024	0.287±.018	0.295±.018	0.515

Table 16: W_1 of each generator over that of the sample, floor removed in quadrature, one standard error over paired replicates; raw: the same ratio at the largest n with the floor left in. Arena C, the unbent sphere and the stronger bend.

C: S^3 **in** $\mathbb{R}^8$**, bend 0**	$n = 203$	$n = 406$	$n = 812$	$n = 1625$	raw
SMOTE, $k = 5$	0.918±.006	0.916±.004	0.924±.006	0.936±.005	0.948
mixing, $K = 2$	0.700±.031	0.668±.026	0.674±.024	0.672±.013	0.745
mixing, $K = 3$	0.736±.029	0.716±.024	0.712±.025	0.720±.011	0.780
rule k, $K = 3$	0.825±.014	0.803±.007	0.797±.015	0.790±.008	0.834
Gaussian kernel	0.893±.003	0.897±.002	0.907±.002	0.928±.001	0.942
PCA kernel	0.711±.009	0.731±.006	0.739±.007	0.761±.004	0.812
oracle chart kernel	0.245±.020	0.284±.018	0.244±.024	0.266±.025	0.497
C: S^3 **in** $\mathbb{R}^{16}$**, bend 1**	$n = 203$	$n = 406$	$n = 812$	$n = 1625$	raw
SMOTE, $k = 5$	0.921±.004	0.917±.001	0.923±.002	0.935±.002	0.947
mixing, $K = 2$	0.752±.018	0.703±.010	0.709±.008	0.699±.008	0.765
mixing, $K = 3$	0.798±.018	0.751±.006	0.757±.007	0.741±.012	0.797
rule k, $K = 3$	0.852±.010	0.812±.005	0.809±.003	0.794±.006	0.837
Gaussian kernel	0.956±.001	0.956±.000	0.962±.000	0.967±.001	0.974
PCA kernel	0.775±.006	0.758±.002	0.770±.001	0.785±.004	0.830
oracle chart kernel	0.282±.026	0.240±.013	0.220±.020	0.237±.048	0.491

Table 17: W_1 of each generator over that of the sample, floor removed in quadrature, one standard error over paired replicates; raw: the same ratio at the largest n with the floor left in. Arena D. †At the estimation floor: on the uniform torus the tuned oracle bandwidth covers the torus and regenerates the population.

D: T^4 in $\mathbb{R}^{16}$, uniform	$n = 90$	$n = 181$	$n = 362$	$n = 724$	raw
SMOTE, $k = 5$	0.899±.002	0.906±.002	0.904±.002	0.905±.002	0.925
mixing, $K = 2$	0.755±.003	0.754±.003	0.748±.003	0.726±.003	0.789
mixing, $K = 3$	0.740±.004	0.755±.004	0.745±.004	0.730±.004	0.792
mixing, $K = 4$	0.778±.004	0.785±.005	0.781±.004	0.768±.004	0.820
rule k, $K = 4$	0.868±.004	0.866±.004	0.837±.004	0.829±.004	0.866
Gaussian kernel	0.903±.001	0.914±.001	0.929±.000	0.930±.000	0.945
PCA kernel	0.766±.002	0.785±.001	0.804±.001	0.841±.001	0.876
oracle chart kernel	0.009±.004†	0.008±.005†	0.011±.007†	0.003±.003†	0.448
D: T^4 in $\mathbb{R}^{16}$, cos	$n = 90$	$n = 181$	$n = 362$	$n = 724$	raw
SMOTE, $k = 5$	0.900±.004	0.899±.004	0.903±.002	0.904±.002	0.924
mixing, $K = 2$	0.772±.002	0.756±.006	0.752±.003	0.733±.004	0.793
mixing, $K = 3$	0.760±.007	0.763±.010	0.761±.004	0.746±.005	0.803
mixing, $K = 4$	0.795±.009	0.793±.010	0.799±.004	0.789±.004	0.835
rule k, $K = 4$	0.880±.006	0.866±.007	0.846±.004	0.839±.004	0.873
Gaussian kernel	0.909±.001	0.920±.001	0.929±.001	0.931±.000	0.945
PCA kernel	0.777±.002	0.792±.001	0.818±.002	0.842±.001	0.876
oracle chart kernel	0.284±.014	0.290±.017	0.306±.014	0.323±.008	0.532

Table 18: Paired head-to-head ratios, floor removed in quadrature, one standard error, and their log-log slope in n. Below one: the first generator is better. Arenas B and C.

B: bumps in $[0,1]^3$	$n=203$	$n=406$	$n=812$	$n=1625$	slope
K=3 / Gauss.	1.312±.085	1.432±.060	1.464±.046	1.495±.040	$+0.048 \pm 0.027$
K=3 / refl.	1.357±.085	1.483±.062	1.516±.049	1.554±.043	$+0.051 \pm 0.027$
K=2 / Gauss.	1.209±.078	1.320±.069	1.371±.046	1.421±.043	$+0.069 \pm 0.029$
K=2 / refl.	1.251±.077	1.367±.072	1.419±.049	1.477±.046	$+0.072 \pm 0.029$
K=2 / K=3	0.922±.006	0.919±.011	0.936±.005	0.950±.005	$+0.015 \pm 0.004$
B: bumps in $[0,1]^4$	$n=90$	$n=181$	$n=362$	$n=724$	slope
K=4 / Gauss.	1.403±.045	1.523±.049	1.540±.049	1.626±.032	$+0.065 \pm 0.017$
K=4 / refl.	1.487±.045	1.621±.054	1.647±.056	1.758±.037	$+0.076 \pm 0.017$
K=2 / Gauss.	1.240±.035	1.319±.029	1.342±.040	1.468±.031	$+0.077 \pm 0.016$
K=2 / refl.	1.314±.035	1.404±.032	1.435±.046	1.586±.034	$+0.087 \pm 0.016$
K=2 / K=4	0.885±.010	0.869±.020	0.872±.013	0.902±.003	$+0.013 \pm 0.005$
B: bumps in $[0,1]^5$	$n=90$	$n=181$	$n=362$		slope
K=5 / Gauss.	1.448±.051	1.475±.058	1.517±.039		$+0.034 \pm 0.031$
K=5 / refl.	1.542±.052	1.586±.063	1.640±.041		$+0.045 \pm 0.030$
K=2 / Gauss.	1.221±.037	1.231±.035	1.266±.041		$+0.026 \pm 0.032$
K=2 / refl.	1.300±.037	1.324±.038	1.369±.043		$+0.037 \pm 0.030$
K=2 / K=5	0.844±.008	0.838±.012	0.834±.009		-0.009 ± 0.011
C: S^3 in $\mathbb{R}^8$, bend 0.3	$n=203$	$n=406$	$n=812$	$n=1625$	slope
K=3 / Gauss.	0.852±.022	0.831±.021	0.813±.012	0.774±.008	-0.050 ± 0.011
K=3 / PCA	1.054±.020	1.009±.021	0.990±.011	0.943±.008	-0.054 ± 0.009
K=3 / oracle	3.573±.342	3.323±.487	2.970±.264	2.672±.150	-0.142 ± 0.051
K=2 / Gauss.	0.810±.022	0.783±.023	0.746±.013	0.723±.008	-0.055 ± 0.012
K=2 / K=3	0.950±.004	0.941±.004	0.917±.006	0.934±.003	-0.008 ± 0.002
K=2 / PCA	1.001±.021	0.950±.023	0.908±.013	0.880±.008	-0.059 ± 0.010
K=2 / oracle	3.397±.329	3.134±.466	2.727±.252	2.496±.141	-0.150 ± 0.052
C: S^3 in $\mathbb{R}^{16}$, bend 0.3	$n=203$	$n=406$	$n=812$	$n=1625$	slope
K=3 / Gauss.	0.804±.016	0.750±.010	0.757±.009	0.739±.012	-0.029 ± 0.011
K=3 / PCA	1.044±.016	0.982±.010	0.982±.011	0.944±.013	-0.039 ± 0.009
K=3 / oracle	2.879±.156	2.852±.145	3.477±.368	3.431±.400	$+0.093 \pm 0.054$
K=2 / Gauss.	0.752±.016	0.705±.010	0.707±.009	0.692±.013	-0.030 ± 0.012
K=2 / K=3	0.935±.005	0.940±.004	0.935±.005	0.937±.007	-0.000 ± 0.004
K=2 / PCA	0.976±.015	0.923±.011	0.918±.011	0.884±.016	-0.041 ± 0.010
K=2 / oracle	2.688±.139	2.681±.140	3.255±.356	3.226±.390	$+0.093 \pm 0.054$
C: S^3 in $\mathbb{R}^{32}$, bend 0.3	$n=203$	$n=406$	$n=812$	$n=1625$	slope
K=3 / Gauss.	0.767±.014	0.743±.014	0.725±.009	0.707±.008	-0.038 ± 0.010
K=3 / PCA	1.025±.017	0.979±.015	0.953±.008	0.917±.008	-0.052 ± 0.008
K=3 / oracle	2.966±.440	2.407±.255	2.533±.129	2.416±.160	-0.058 ± 0.065
K=2 / Gauss.	0.723±.012	0.691±.012	0.680±.009	0.658±.008	-0.043 ± 0.009
K=2 / K=3	0.943±.006	0.931±.005	0.938±.004	0.932±.005	-0.004 ± 0.004
K=2 / PCA	0.967±.015	0.911±.013	0.894±.009	0.855±.008	-0.055 ± 0.008
K=2 / oracle	2.802±.423	2.238±.232	2.377±.123	2.252±.150	-0.060 ± 0.065
C: S^3 in $\mathbb{R}^8$, bend 0	$n=203$	$n=406$	$n=812$	$n=1625$	slope
K=3 / Gauss.	0.824±.030	0.799±.025	0.785±.026	0.776±.011	-0.026 ± 0.016
K=3 / PCA	1.034±.028	0.980±.025	0.963±.025	0.945±.013	-0.038 ± 0.013
K=3 / oracle	3.098±.410	2.557±.176	3.064±.494	2.779±.275	$+0.004 \pm 0.070$
K=2 / Gauss.	0.783±.032	0.744±.027	0.743±.024	0.724±.013	-0.032 ± 0.019
K=2 / K=3	0.950±.006	0.931±.007	0.947±.007	0.934±.005	-0.006 ± 0.004
K=2 / PCA	0.982±.032	0.913±.027	0.911±.024	0.883±.014	-0.044 ± 0.015
K=2 / oracle	2.949±.410	2.380±.157	2.902±.472	2.595±.256	$+0.006 \pm 0.071$
C: S^3 in $\mathbb{R}^{16}$, bend 1	$n=203$	$n=406$	$n=812$	$n=1625$	slope
K=3 / Gauss.	0.834±.018	0.786±.005	0.787±.007	0.766±.012	-0.022 ± 0.010
K=3 / PCA	1.029±.021	0.991±.005	0.984±.008	0.944±.012	-0.031 ± 0.008
K=3 / oracle	2.896±.275	3.165±.184	3.531±.334	3.411±.485	$+0.102 \pm 0.072$
K=2 / Gauss.	0.787±.018	0.736±.010	0.736±.009	0.722±.008	-0.028 ± 0.010
K=2 / K=3	0.943±.007	0.936±.008	0.935±.006	0.943±.006	$+0.000 \pm 0.004$
K=2 / PCA	0.971±.019	0.928±.012	0.921±.011	0.890±.008	-0.036 ± 0.008
K=2 / oracle	2.727±.243	2.964±.184	3.303±.311	3.222±.464	$+0.103 \pm 0.070$

Table 19: Paired head-to-head ratios, floor removed in quadrature, one standard error, and their log-log slope in n. Below one: the first generator is better. Arena D.

D: T^4 in $\mathbb{R}^{16}$, uniform	$n = 90$	$n = 181$	$n = 362$	$n = 724$	slope
K=4 / Gauss.	0.862±.004	0.859±.005	0.841±.004	0.826±.004	-0.020 ± 0.003
K=4 / PCA	1.015±.003	1.000±.005	0.971±.004	0.913±.004	-0.046 ± 0.002
K=2 / Gauss.	0.836±.002	0.826±.003	0.806±.003	0.780±.003	-0.032 ± 0.002
K=2 / K=4	0.970±.002	0.962±.003	0.958±.002	0.945±.002	-0.012 ± 0.001
K=2 / K=3	1.020±.002	0.999±.002	1.004±.002	0.995±.002	-0.009 ± 0.001
K=3 / Gauss.	0.820±.004	0.826±.004	0.803±.004	0.784±.004	-0.022 ± 0.003
K=3 / PCA	0.966±.003	0.963±.004	0.927±.004	0.867±.004	-0.048 ± 0.003
K=3 / K=4	0.951±.001	0.963±.002	0.954±.001	0.950±.001	-0.001 ± 0.001
K=2 / PCA	0.985±.002	0.962±.002	0.930±.003	0.863±.003	-0.055 ± 0.002
D: T^4 in $\mathbb{R}^{16}$, cos	$n = 90$	$n = 181$	$n = 362$	$n = 724$	slope
K=4 / Gauss.	0.874±.010	0.863±.011	0.861±.004	0.847±.005	-0.016 ± 0.005
K=4 / PCA	1.023±.012	1.001±.014	0.977±.005	0.937±.005	-0.045 ± 0.006
K=4 / oracle	2.859±.187	2.831±.228	2.651±.126	2.457±.076	-0.080 ± 0.032
K=2 / Gauss.	0.849±.002	0.822±.006	0.810±.004	0.787±.004	-0.035 ± 0.002
K=2 / K=4	0.972±.009	0.953±.006	0.941±.002	0.929±.001	-0.019 ± 0.003
K=2 / K=3	1.016±.007	0.991±.005	0.988±.001	0.982±.002	-0.012 ± 0.003
K=3 / Gauss.	0.835±.007	0.829±.011	0.820±.004	0.801±.006	-0.020 ± 0.005
K=3 / PCA	0.978±.009	0.962±.013	0.930±.005	0.887±.006	-0.047 ± 0.005
K=3 / K=4	0.956±.003	0.961±.001	0.952±.002	0.946±.002	-0.009 ± 0.001
K=2 / PCA	0.994±.003	0.954±.008	0.919±.005	0.870±.005	-0.061 ± 0.003
K=2 / oracle	2.769±.160	2.690±.203	2.495±.121	2.283±.071	-0.098 ± 0.030

Table 20: Weighted log-log slope in n of each ratio to the sample (floor removed in quadrature), one standard error, and below it in brackets the slope the exponents of Theorem A predict (runs named by arena and d; m in place of d on a manifold, where Theorem A is not proved; zero for a kernel whose off-manifold bias is first order). Sample: the slope of the sample's own W_1, predicted $-1/m$. The $K = 3$ column is filled where $3 < m$.

run	$K = 2$	$K = 3$	$K = m$	rule, $K = m$	Gauss.	refl./PCA	sample
B3	$+0.027 \pm .020$		$+0.006 \pm .016$	$-0.005 \pm .015$	$-0.050 \pm .015$	$-0.052 \pm .015$	-0.382
	[−0.067]		[−0.067]	[−0.067]	[−0.067]	[−0.067]	[−0.333]
B4	$+0.020 \pm .011$		$+0.007 \pm .012$	$-0.013 \pm .009$	$-0.057 \pm .012$	$-0.066 \pm .013$	-0.303
	[−0.036]		[−0.083]	[−0.083]	[−0.083]	[−0.083]	[−0.250]
B5	$+0.005 \pm .019$		$+0.017 \pm .018$	$-0.024 \pm .011$	$-0.019 \pm .018$	$-0.029 \pm .017$	-0.245
	[−0.022]		[−0.086]	[−0.086]	[−0.086]	[−0.086]	[−0.200]
C8	$-0.036 \pm .013$		$-0.031 \pm .012$	$-0.029 \pm .007$	$+0.020 \pm .001$	$+0.023 \pm .004$	-0.355
	[−0.067]		[−0.067]	[−0.067]	[+0.000]	[+0.000]	[−0.333]
C16	$-0.023 \pm .013$		$-0.021 \pm .012$	$-0.026 \pm .006$	$+0.008 \pm .000$	$+0.021 \pm .003$	-0.361
	[−0.067]		[−0.067]	[−0.067]	[+0.000]	[+0.000]	[−0.333]
C32	$-0.041 \pm .010$		$-0.037 \pm .010$	$-0.034 \pm .006$	$+0.001 \pm .000$	$+0.014 \pm .004$	-0.361
	[−0.067]		[−0.067]	[−0.067]	[+0.000]	[+0.000]	[−0.333]
C8, unbent	$-0.011 \pm .020$		$-0.005 \pm .017$	$-0.016 \pm .008$	$+0.021 \pm .001$	$+0.031 \pm .005$	-0.350
	[−0.067]		[−0.067]	[−0.067]	[+0.000]	[+0.000]	[−0.333]
C16, bend 1	$-0.021 \pm .011$		$-0.015 \pm .010$	$-0.024 \pm .005$	$+0.007 \pm .000$	$+0.014 \pm .003$	-0.363
	[−0.067]		[−0.067]	[−0.067]	[+0.000]	[+0.000]	[−0.333]
D, unif.	$-0.016 \pm .003$	$-0.008 \pm .004$	$-0.005 \pm .003$	$-0.024 \pm .003$	$+0.014 \pm .000$	$+0.046 \pm .001$	-0.259
	[−0.036]	[−0.083]	[−0.083]	[−0.083]	[+0.000]	[+0.000]	[−0.250]
D, cos.	$-0.023 \pm .002$	$-0.008 \pm .005$	$-0.005 \pm .006$	$-0.024 \pm .004$	$+0.010 \pm .000$	$+0.041 \pm .001$	-0.259
	[−0.036]	[−0.083]	[−0.083]	[−0.083]	[+0.000]	[+0.000]	[−0.250]

Table 21: Checks of the predictions in arenas C and D (Section I.2): verdict at two standard errors under the three floor removals (quadrature is primary; calibrated is n/a where no exponent calibrates the sample's decay), and the numbers behind the primary verdict.

check	quadr.	none	calibr.	primary numbers
C1 d=16 bend=0.3	PASS	INC.	INC.	slope(Gauss./sample) +0.008 vs slope(K=3/sample) −0.021 (diff +0.029 ± 0.012)
C1 d=32 bend=0.3	PASS	INC.	PASS	slope(Gauss./sample) +0.001 vs slope(K=3/sample) −0.037 (diff +0.038 ± 0.010)
C1 d=8 bend=0.0	INC.	INC.	INC.	slope(Gauss./sample) +0.021 vs slope(K=3/sample) −0.005 (diff +0.026 ± 0.017)
C1 d=8 bend=0.3	PASS	PASS	PASS	slope(Gauss./sample) +0.020 vs slope(K=3/sample) −0.031 (diff +0.051 ± 0.012)
C2 d=16 bend=0.3	INC.	INC.	INC.	K=3/sample raw max 0.781; slope −0.021 ± 0.012
C2 d=32 bend=0.3	PASS	INC.	INC.	K=3/sample raw max 0.765; slope −0.037 ± 0.010
C2 d=8 bend=0.0	INC.	INC.	INC.	K=3/sample raw max 0.780; slope −0.005 ± 0.017
C2 d=8 bend=0.3	PASS	INC.	INC.	K=3/sample raw max 0.780; slope −0.031 ± 0.012
C3 d=16 bend=0.3	PASS	PASS	PASS	K=3/Gauss. per n: 0.804, 0.750, 0.757, 0.739
C3 d=32 bend=0.3	PASS	PASS	PASS	K=3/Gauss. per n: 0.767, 0.743, 0.725, 0.707
C3 d=8 bend=0.0	PASS	PASS	PASS	K=3/Gauss. per n: 0.824, 0.799, 0.785, 0.776
C3 d=8 bend=0.3	PASS	PASS	PASS	K=3/Gauss. per n: 0.852, 0.831, 0.813, 0.774
C4 d=16 bend=0.3	PASS	PASS	PASS	slope(PCA/sample) +0.021 vs slope(K=3/sample) −0.021 (diff +0.042 ± 0.012)
C4 d=32 bend=0.3	PASS	PASS	PASS	slope(PCA/sample) +0.014 vs slope(K=3/sample) −0.037 (diff +0.051 ± 0.011)
C4 d=8 bend=0.0	PASS	INC.	INC.	slope(PCA/sample) +0.031 vs slope(K=3/sample) −0.005 (diff +0.037 ± 0.018)
C4 d=8 bend=0.3	PASS	PASS	PASS	slope(PCA/sample) +0.023 vs slope(K=3/sample) −0.031 (diff +0.054 ± 0.013)
D1a d=16 cos	PASS	PASS	PASS	K=4/Gauss. per n: 0.874, 0.863, 0.861, 0.847
D1a d=16 uniform	PASS	PASS	PASS	K=4/Gauss. per n: 0.862, 0.859, 0.841, 0.826
D1b d=16 cos	INC.	INC.	INC.	K=4/PCA per n: 1.023, 1.001, 0.977, 0.937
D1b d=16 uniform	FAIL	FAIL	FAIL	K=4/PCA per n: 1.015, 1.000, 0.971, 0.913
D2 d=16 uniform K=4	FAIL	FAIL	FAIL	K=2/K=4 per n: 0.970, 0.962, 0.958, 0.945 (> 1: K=4 better)
D3 m=4 d=16	PASS	PASS	PASS	mean log(K=2/K=4): uniform −0.042 ± 0.001, cos −0.053 ± 0.003; diff +0.011 ± 0.003
D4 d=16 cos	PASS	PASS	PASS	K=4/sample raw per n: 0.813, 0.818, 0.832, 0.835
D4 d=16 uniform	PASS	PASS	PASS	K=4/sample raw per n: 0.798, 0.810, 0.817, 0.820

Table 22: Further checks: arena B, the bend-1 run of C, the dependence on d in C, and $K = 3$ in D (Section I.2): verdict at two standard errors under the three floor removals (quadrature is primary; calibrated is n/a where no exponent calibrates the sample's decay), and the numbers behind the primary verdict.

check	quadr.	none	calibr.	primary numbers
B-level d=3 K=2/Gauss.	kernel	kernel	n/a	per n: 1.209, 1.320, 1.371, 1.421; slope +0.069 ± 0.029
B-level d=3 K=2/refl.	kernel	kernel	n/a	per n: 1.251, 1.367, 1.419, 1.477; slope +0.072 ± 0.029
B-level d=3 K=3/Gauss.	kernel	kernel	n/a	per n: 1.312, 1.432, 1.464, 1.495; slope +0.048 ± 0.027
B-level d=3 K=3/refl.	kernel	kernel	n/a	per n: 1.357, 1.483, 1.516, 1.554; slope +0.051 ± 0.027
B-level d=4 K=2/Gauss.	kernel	kernel	n/a	per n: 1.240, 1.319, 1.342, 1.468; slope +0.077 ± 0.016
B-level d=4 K=2/refl.	kernel	kernel	n/a	per n: 1.314, 1.404, 1.435, 1.586; slope +0.087 ± 0.016
B-level d=4 K=4/Gauss.	kernel	kernel	n/a	per n: 1.403, 1.523, 1.540, 1.626; slope +0.065 ± 0.017
B-level d=4 K=4/refl.	kernel	kernel	n/a	per n: 1.487, 1.621, 1.647, 1.758; slope +0.076 ± 0.017
B-level d=5 K=2/Gauss.	kernel	kernel	n/a	per n: 1.221, 1.231, 1.266; slope +0.026 ± 0.032
B-level d=5 K=2/refl.	kernel	kernel	n/a	per n: 1.300, 1.324, 1.369; slope +0.037 ± 0.030
B-level d=5 K=5/Gauss.	kernel	kernel	n/a	per n: 1.448, 1.475, 1.517; slope +0.034 ± 0.031
B-level d=5 K=5/refl.	kernel	kernel	n/a	per n: 1.542, 1.586, 1.640; slope +0.045 ± 0.030
B-rate d=3 refl./sample	PASS	FAIL	n/a	slope −0.052 ± 0.015, predicted −0.067
B-rate d=3 K=3/sample	INC.	INC.	n/a	slope +0.006 ± 0.016, predicted −0.067
B-rate d=4 refl./sample	PASS	FAIL	n/a	slope −0.066 ± 0.013, predicted −0.083
B-rate d=4 K=4/sample	INC.	INC.	n/a	slope +0.007 ± 0.012, predicted −0.083
B-rate d=5 refl./sample	INC.	FAIL	n/a	slope −0.029 ± 0.017, predicted −0.086
B-rate d=5 K=5/sample	INC.	INC.	n/a	slope +0.017 ± 0.018, predicted −0.086
C1 d=16 bend=1.0	PASS	INC.	INC.	slope(Gauss./sample) +0.007 vs slope(K=3/sample) −0.015 (diff +0.022 ± 0.010)
C1b m=3 bend=0.3	PASS	PASS	PASS	Gauss./sample at n=1625 over d=[8, 16, 32]: 0.927, 0.966, 0.982
C2 d=16 bend=1.0	INC.	INC.	INC.	K=3/sample raw max 0.809; slope −0.015 ± 0.010
C3 d=16 bend=1.0	PASS	PASS	PASS	K=3/Gauss. per n: 0.834, 0.786, 0.787, 0.766
C3b m=3 bend=0.3	PASS	PASS	PASS	K=3/Gauss. at n=1625 over d: 0.774, 0.739, 0.707
C4 d=16 bend=1.0	PASS	PASS	PASS	slope(PCA/sample) +0.014 vs slope(K=3/sample) −0.015 (diff +0.028 ± 0.011)
D2 d=16 cos K=4	FAIL	FAIL	FAIL	K=2/K=4 per n: 0.972, 0.953, 0.941, 0.929 (> 1: K=4 better)
D2' d=16 cos K=3	FAIL	FAIL	FAIL	K=2/K=3 per n: 1.016, 0.991, 0.988, 0.982 (> 1: K=3 better)
D2' d=16 uniform K=3	FAIL	FAIL	FAIL	K=2/K=3 per n: 1.020, 0.999, 1.004, 0.995 (> 1: K=3 better)

Table 23: Selected scale per n on the tuning replicates: k for mixing, the bandwidth for a kernel, and in brackets its multiple of the grid centre (the rule (2) for mixing; for a kernel, the Gaussian matching a pair's spread at the rule's radius). ^{e}: on a grid edge ($k = n - 1$ cannot be extended). Only the n that were evaluated are shown.

run	method	selected scale [multiple] at increasing n			
B, $d = 3$	mixing, $K = 2$	36 [4]	48 [4]	30 [2]	40 [2]
	mixing, $K = 3$	18 [2]	12 [1]	15 [1]	20 [1]
	Gaussian kernel	0.0567 [1]	0.0495 [1]	0.0423 [1]	0.037 [1]
	reflected kernel	0.0567 [1]	0.0495 [1]	0.0423 [1]	0.037 [1]
B, $d = 4$	mixing, $K = 2$	56 [8]	80 [8]	104 [8]	68 [4]
	mixing, $K = 4$	20 [4]	12 [2]	16 [2]	18 [2]
	Gaussian kernel	0.0835 [1]	0.0767 [1]	0.0688 [1]	0.0619 [1]
	reflected kernel	0.0835 [1]	0.0767 [1]	0.0688 [1]	0.0619 [1]
B, $d = 5$	mixing, $K = 2$	89 [11.1]^{e}	180 [16.4]^{e}	112 [8]	
	mixing, $K = 5$	16 [4]	20 [4]	12 [2]	
	Gaussian kernel	0.0965 [1]	0.0894 [1]	0.0817 [1]	
	reflected kernel	0.0965 [1]	0.0894 [1]	0.0817 [1]	
C, $d = 16$, bend 1	mixing, $K = 2$	72 [8]	96 [8]	120 [8]	160 [8]
	mixing, $K = 3$	36 [4]	48 [4]	60 [4]	80 [4]
	Gaussian kernel	0.0192 [0.125]	0.0167 [0.125]	0.0143 [0.125]	0.00625 [0.0625]
	PCA kernel	0.153 [1]	0.0669 [0.5]	0.0572 [0.5]	0.05 [0.5]
	oracle chart kernel	0.613 [4]	0.535 [4]	0.458 [4]	0.4 [4]
C, $d = 16$	mixing, $K = 2$	72 [8]	96 [8]	120 [8]	160 [8]
	mixing, $K = 3$	36 [4]	48 [4]	60 [4]	80 [4]
	Gaussian kernel	0.0192 [0.125]	0.0167 [0.125]	0.00715 [0.0625]	0.00625 [0.0625]
	PCA kernel	0.153 [1]	0.0669 [0.5]	0.0572 [0.5]	0.05 [0.5]
	oracle chart kernel	0.613 [4]	0.535 [4]	0.458 [4]	0.4 [4]
C, $d = 32$	mixing, $K = 2$	72 [8]	96 [8]	120 [8]	160 [8]
	mixing, $K = 3$	36 [4]	48 [4]	60 [4]	80 [4]
	Gaussian kernel	0.00958 [0.0625]	0.00837 [0.0625]	0.00358 [0.0312]	0.00312 [0.0312]
	PCA kernel	0.153 [1]	0.0669 [0.5]	0.0572 [0.5]	0.05 [0.5]
	oracle chart kernel	0.613 [4]	0.535 [4]	0.458 [4]	0.4 [4]
C, $d = 8$, unbent	mixing, $K = 2$	72 [8]	96 [8]	120 [8]	160 [8]
	mixing, $K = 3$	36 [4]	48 [4]	60 [4]	80 [4]
	Gaussian kernel	0.0383 [0.25]	0.0335 [0.25]	0.0286 [0.25]	0.0125 [0.125]
	PCA kernel	0.153 [1]	0.0669 [0.5]	0.0572 [0.5]	0.05 [0.5]
	oracle chart kernel	0.613 [4]	0.535 [4]	0.458 [4]	0.4 [4]
C, $d = 8$	mixing, $K = 2$	72 [8]	96 [8]	120 [8]	160 [8]
	mixing, $K = 3$	36 [4]	48 [4]	30 [2]	80 [4]
	Gaussian kernel	0.0383 [0.25]	0.0335 [0.25]	0.0286 [0.25]	0.0125 [0.125]
	PCA kernel	0.153 [1]	0.0669 [0.5]	0.0572 [0.5]	0.05 [0.5]
	oracle chart kernel	0.613 [4]	0.535 [4]	0.458 [4]	0.4 [4]
D, cosine	mixing, $K = 2$	89 [12.7]^{e}	80 [8]	52 [4]	68 [4]
	mixing, $K = 3$	40 [8]	24 [4]	32 [4]	36 [4]
	mixing, $K = 4$	20 [4]	24 [4]	16 [2]	18 [2]
	Gaussian kernel	0.0209 [0.25]	0.0192 [0.25]	0.00861 [0.125]	0.00774 [0.125]
	PCA kernel	0.0418 [0.5]	0.0383 [0.5]	0.0344 [0.5]	0.0155 [0.25]
	oracle chart kernel	0.167 [2]	0.153 [2]	0.138 [2]	0.124 [2]
D, uniform	mixing, $K = 2$	56 [8]	80 [8]	104 [8]	68 [4]
	mixing, $K = 3$	40 [8]	48 [8]	32 [4]	36 [4]
	mixing, $K = 4$	40 [8]	24 [4]	32 [4]	36 [4]
	Gaussian kernel	0.0209 [0.25]	0.0192 [0.25]	0.00861 [0.125]	0.00774 [0.125]
	PCA kernel	0.0418 [0.5]	0.0383 [0.5]	0.0344 [0.5]	0.0155 [0.25]
	oracle chart kernel	0.668 [8]	1.23 [16]	0.551 [8]	0.99 [16]

Table 24: Neighbour check at full N: change in W_1 (percent, one standard error) when the selected scale is halved or doubled, per n. Positive: the neighbour is worse, so the selection sits at a minimum.

B: bumps in $[0,1]^4$		$n = 90$	$n = 181$	$n = 362$	$n = 724$
mixing, $K = 2$	$\times 0.5$	$+1.9 \pm 1.5$	$+1.3 \pm 0.1$	$+0.6 \pm 1.1$	-0.0 ± 1.0
mixing, $K = 2$	$\times 2$	$+0.9 \pm 2.1$	$+4.9 \pm 1.1$	$+7.0 \pm 1.8$	$+0.6 \pm 0.2$
mixing, $K = 4$	$\times 0.5$	-1.7 ± 0.1	$+2.9 \pm 0.0$	$+2.3 \pm 1.1$	$+0.4 \pm 1.0$
mixing, $K = 4$	$\times 2$	$+8.5 \pm 1.1$	$+0.4 \pm 0.4$	$+3.0 \pm 0.9$	$+5.1 \pm 0.6$
Gaussian kernel	$\times 0.5$	$+18.9 \pm 0.4$	$+19.2 \pm 0.1$	$+13.7 \pm 0.9$	$+13.1 \pm 0.6$
Gaussian kernel	$\times 2$	$+30.6 \pm 0.6$	$+37.6 \pm 2.7$	$+35.6 \pm 0.4$	$+41.0 \pm 1.3$
reflected kernel	$\times 0.5$	$+23.0 \pm 0.7$	$+23.7 \pm 0.4$	$+17.7 \pm 1.1$	$+17.3 \pm 0.5$
reflected kernel	$\times 2$	$+5.6 \pm 0.8$	$+11.0 \pm 2.7$	$+13.1 \pm 0.7$	$+19.4 \pm 1.7$
C: S^3 **in** $\mathbb{R}^{16}$**, bend 0.3**		$n = 203$	$n = 406$	$n = 812$	$n = 1625$
mixing, $K = 2$	$\times 0.5$	$+2.1 \pm 0.5$	$+1.0 \pm 1.0$	$+1.6 \pm 0.2$	$+1.4 \pm 0.7$
mixing, $K = 2$	$\times 2$	$+8.1 \pm 2.1$	$+7.5 \pm 0.9$	$+6.7 \pm 1.5$	$+6.1 \pm 1.7$
mixing, $K = 3$	$\times 0.5$	$+2.3 \pm 2.2$	$+1.8 \pm 1.2$	$+0.9 \pm 0.6$	$+1.1 \pm 0.6$
mixing, $K = 3$	$\times 2$	$+6.7 \pm 0.4$	$+7.7 \pm 1.5$	$+6.5 \pm 0.3$	$+6.3 \pm 1.8$
Gaussian kernel	$\times 0.5$	$+1.1 \pm 0.0$	$+0.7 \pm 0.0$	$+1.0 \pm 0.0$	$+0.8 \pm 0.0$
Gaussian kernel	$\times 2$	$+3.8 \pm 0.0$	$+5.6 \pm 0.0$	-0.1 ± 0.0	$+0.7 \pm 0.0$
PCA kernel	$\times 0.5$	$+1.8 \pm 0.1$	$+11.0 \pm 0.0$	$+8.8 \pm 0.0$	$+6.4 \pm 0.1$
PCA kernel	$\times 2$	$+39.9 \pm 0.8$	$+2.5 \pm 0.2$	$+5.6 \pm 0.1$	$+9.6 \pm 0.6$
oracle chart kernel	$\times 0.5$	$+18.6 \pm 5.7$	$+11.4 \pm 2.2$	$+11.5 \pm 0.2$	$+8.8 \pm 1.2$
oracle chart kernel	$\times 2$	$+88.3 \pm 6.1$	$+84.3 \pm 1.9$	$+81.4 \pm 1.6$	$+56.9 \pm 0.2$

J SUMMARY TABLES

Table 25 compares the methods discussed in the paper, and Tables 26 and 27 collect the constants and exponents for which the paper gives both a proved or predicted value and a measured one. The per-dimension measurements behind several rows are in Table 3.

Table 25: Comparison of the methods discussed in the paper, for $d \geq 3$ and a density satisfying Assumption 1. Rates are for $\mathbb{E}W_1$ between the synthetic distribution and μ_0. The empirical measure is $K = 1$, Gaussian smoothing is $\hat{\mu}_n * \mathcal{N}_\sigma$, global mixup mixes pairs with no locality, and both SMOTE columns use the k-NN rule. The measured rows give W_1 over that of the sample, and the notes below explain (a)–(d).

	Empirical measure	Gaussian smoothing	Global mixup	SMOTE $K = 2$	SMOTE $K \geq d - 1$
Tuning	–	$\sigma \asymp n^{-1/(d+2)}$	–	$k \asymp n^{\frac{d-1}{2d-1}}$	$k \asymp n^{\frac{2}{d+2}}$
Directions smoothed	0	d	1	1	$\min(K-1, d-1)$
Bias $W_1(\bar{\nu}, \mu_0)$	0	$O(\sigma^2)$	$\Theta(1)$	$\Theta(h_\star^2)$	$\Theta(h_\star^2)$
Bias constant	–	–	–	$c_2(1)(1 + \frac{1}{d})$	$c_K(1)(1 + \frac{2}{dK})$
Rate	$n^{-1/d}$	$n^{-2/(d+2)}$ (a)	none (b)	$n^{-2/(2d-1)}$ (d)	$n^{-2/(d+2)}$ (c)
Distance to a manifold	0	$\geq \sigma\sqrt{d-m}/5$	$\leq D^2/\tau$	$\leq h_\star^2/\tau$	$\leq h_\star^2/\tau$
W_1/sample, bumps	1	0.52–0.56	–	0.68–0.79	0.82–0.84
W_1/sample, S^3	1	0.93–0.98	–	0.65–0.67	0.69–0.72
In the convex hull	yes	no	yes	yes	yes
Mixes labels	–	no	yes	yes	yes
Scale set by	–	σ	–	k	k
Results	classical	Lem. D.5, Prop. 5.1	Cor. A.13	Thm. A	Thms. A, B

Notes to Table 25. (a) On the torus (Lemma D.5), where the bias is at most $\frac{1}{2}\sigma^2\|\nabla f\|_{L^1}$. (b) Global mixup does not converge to μ_0. It converges at $n^{-1/(d-1)}$ to the distribution $\mathcal{S}_K(\mu_0)$ of a random convex combination of independent draws (Corollary A.13). (c) With a factor $\log n$ at $K = d - 1$. This rate is optimal among local convex mixing with a bias of order h^2, by the lower bound of Theorem B, proved on the torus for a non-flat density. The bias constants are those of Proposition 4.1, $c_K(1) = \frac{K-1}{(K+1)(d+2)}$, multiplied by the k-NN factor $1 + 2/(dK)$ of Lemma E.5. The ε-ball rule has the rates of the k-NN rule with $\varepsilon \propto n^{-1/(2d-\bar{\kappa})}$ and no such factor (Remark E.1). For $d = 3$ pairs are the case $K = d - 1$, with rate $n^{-2/5}\log n$. Arities $3 \leq K \leq d - 2$ give $n^{-2/(2d-K+1)}$ with $k \asymp n^{(d-K+1)/(2d-K+1)}$ (Theorem A). The measured rows give W_1 over that of the sample at the largest n in arenas B ($d = 3, 4, 5$) and C ($d = 8, 16, 32$) of Section I.2, with k and the bandwidth tuned and $K = m$ in the last column (Table 12). The Gaussian kernel's scale σ is in the units of the features, whereas k is a count of neighbours. Whatever the generator, a training set of $N \asymp n$ synthetic points, as in SMOTE to class parity at a bounded imbalance ratio, does not beat $n^{-1/d}$ (Corollary 3.1). Over the C^2 class of Assumption 1 the minimax rate is $n^{-3/(d+4)}$. It is attained by estimators with signed kernels (Weed & Berthet, 2019; Niles-Weed & Berthet, 2022), which convex mixing with a fixed or k-NN radius cannot produce, while a radius growing like $f^{K/2}$ removes the interior h^2 term (Remark D.6). (d) The proved rate. Under Assumption 3, pairs attain $n^{-2/5}$ at $d = 3$ and $n^{-1/(d-1)}$ for $d \geq 4$, with a factor $\log n$ at $d = 4$, and every $K > d/2$ attains $n^{-2/(d+2)}$ (Proposition E.10).

Table 26: Constants: proved or predicted values against measured ones.

Quantity	Proved or predicted	Measured	Source
Mixing bias			
$c_K(1)$, fixed radius, $K = 2$, $d = 3, \dots, 6$	.0667, .0556, .0476, .0417	.0676, .0553, .0475, .0417 (s.e. 3–6 $\times 10^{-4}$)	Tab. 3
$c_K(1)$, fixed radius, $K = d$, $d = 3, \dots, 6$	.1000, .1000, .0952, .0893	.0997, .0996, .0950, .0898	Tab. 3
$c_K(1)$, $K = 2, \dots, 12$, $d = 4, 6$	closed form	within 1.5%	Fig. 1B
$c_K(\alpha)$, $d = 4$, $K = 2, 3$, $\alpha \in [0.2, 3]$ and $\alpha \to 0$	closed form	within 2%	Rem. D.3
Share of neighbour selection $c_K^{\mathrm{nn}}(0)/c_K^{\mathrm{nn}}(1)$, k-NN, $d = 4$	0.900 ($K = 2$), 0.857 ($K = 3$)	0.906 ± 0.004, 0.860 ± 0.004	Rem. D.3
k-NN bias over (60), $d = 1$	1	0.94–0.99 (fixed mass q); within 1% (k-NN rule)	Lem. E.5
β_1/h^2, circle, $K = 2$, $f = 1 + a\cos 2\pi x$	$4a/9$	ratio 0.97–0.99 at $h = 0.05$; within 0.5% at $h = 0.02$	Rem. D.4
β_2/h^2, circle, $K = 2$, $a = 0.6$	0.312 (linearised)	0.29, 0.31 at $h = 0.1, 0.05$	Rem. F.2
Error ratio at equal $\bar{\kappa}$: pairs over $K = 3$ at $m = 3$; $K = 3$ over $K = 4$ at $m = 4$	0.938; 0.952 (k-NN bias; 0.922, 0.941 with a fixed radius; heuristic, post hoc)	0.92–0.95, mean 0.937; 0.95–0.96, mean 0.954	App. I.2
Fluctuation			
Cube constant C_Q, $d = 4, 5, 10$	$\sqrt{d}$: 2.00, 2.24, 3.16	effective, at most 1.50, 1.32, 1.37	App. G
Crossover C_Q^{2d}, $d = 5$	3125	at most 16	App. G
Variance factor b_K, pairs, $d = 3, \dots, 6$	at most $K^2 = 4$	1.57–1.82	Tab. 3
$\sqrt{b_K}$, $d = 1$, $K = 2, 3, 5$	at most K	1.47, 1.86, 2.60 (local); 0.78–0.85 (global)	App. A.4
$C_{2,d}$ of Assumption 2, $d = 3, \dots, 6$	at most $\frac{2d\sqrt{d}}{d-1}$, i.e. 5.2–5.9 (fixed radius)	1.11, 0.97, 0.87, 0.81 (k-NN pairs)	Prop. C.3
$\mathbb{E}[1/J]$, normalised inverse simplex volume, $d = 4$, $K = 2, 3, 4$	finite for $K \le d$, infinite at $K = d + 1$	1.33, 2.26, 6.04; grows with n at $K = d + 1$	App. C
Overlap nS_π, uniform base, k-NN graph	$1 + (\frac{K-1}{K})^2 \operatorname{Var}(T) = O(1)$	1.1–1.4	Prop. G.1
Manifold (unit spheres, reach $\tau = 1$)			
Worst mixing defect $\operatorname{dist}(Y, \mathcal{M})/h_\star^2$	at most 1	0.47–0.49	Prop. 5.1
Mean mixing defect on S^2, $/h_\star^2$	$1/24 \approx 0.0417$	0.0413–0.0424	App. H
Kernel displacement $/\sigma\sqrt{d-m}$	at least 0.2	0.88–0.97	Prop. 5.1
Kernel normal displacement on MNIST, $/h_\star$	$\sqrt{772/784} = 0.99$	0.99–1.02	App. H

Notes. The proved values of C_Q, b_K and $C_{2,d}$ are upper bounds, and the measured values lie well below them; the measured C_Q are themselves upper bounds, read off gains observed at small k. $C_{2,d}$ is bounded for a partner drawn uniformly from a ball and measured for k-NN pairs normalised at the k-NN scale, so these two columns compare orders of magnitude only.

Table 27: Exponents: proved or predicted values against measured ones.

Quantity	Proved or predicted	Measured	Source
β_1 in h, circle	2	1.91–1.95, $h \in [0.02, 0.14]$	Rem. D.4
β_2 in h, uniform distribution on $[0, 1]$	$3/2$ (heuristic)	1.49	Rem. F.2
Pair small-ball exponent, $\mathrm{Beta}(\alpha, \alpha)$ weights, $\alpha = 0.15, 0.25, 0.35, 0.5, 0.8$	$\min\{2\alpha, 1\}$, with a logarithm at $\alpha = \frac{1}{2}$	0.30, 0.50, 0.69, 0.89, 1.00	Rem. D.3
Small-ball exponent for $K > d$	$d - 1$ proved	2.33–2.76 ($d = 3$); 3.12–3.74 ($d = 4$)	Lem. E.2
Two draws from one base, k-NN with $k = 24$ (96): pairs, $d = 3$–6; $K = 3$, $d = 5, 6$	$\kappa_0 = 2$, 4 proved, with terms of weight k^{-j}	1.59–1.62 (1.86 at $d = 4$); 2.94, 2.99 (3.40, 3.49)	Lem. E.9
Pool, anchored at the replaced point, $k = 24$ (96): same cases	$K - 1$ proved; $\kappa_0 = 2$, 4 assumed	1.88–2.00 (1.98 at $d = 4$); 3.54, 3.67 (3.75, 3.84)	Ass. 3
Pool, other anchors, $k = 24$ (96): same cases	$K - 1$ proved; $\kappa_0 = 2$, 4 assumed	2.20–2.82 (3.41 at $d = 4$); 3.99, 3.81 (4.60, 4.25)	Ass. 3
Pool exponent, fixed radius	$K - 1$ proved; κ_0 assumed	1.87, 2.09 (pairs, $d = 3, 5$); 3.7, 3.9 ($K = 3$, $d = 5, 6$)	Lem. E.4
Tuned-rate slope of the synthetic level, $d = 3, 4, 5$	-0.400, -0.333, -0.286	-0.388, -0.310, -0.280	Tab. 11
Slope of the ratio to the empirical level, $d = 3, 4, 5$	-0.067, -0.083, -0.086	-0.032, -0.034, -0.049	Tab. 11
Slope of tuned mixing over the sample, arenas B–D	-0.022 to -0.086	-0.041 to $+0.027$	Tab. 20

Notes. The small-ball and pool exponents are proved as lower bounds, and the measured exponents exceed them. The two-draw slopes fall short of κ_0 at $k = 24$ by the terms of (64) with shared partners, which fade as k grows. The tuned-rate slopes are measured with the estimation floor removed in quadrature; Table 11 gives the alternatives, and the slope of the ratio is negative under each of them.