

How Many Samples Does a Generative Evaluation Need?

Intrinsic Dimension and the Design of Distributional Comparisons

Ömer Kahyaoglu

kahyaoglu20@itu.edu.tr
math.kahyaoglu@gmail.com

*Department of Mathematical Engineering
Faculty of Science and Letters
Istanbul Technical University
Istanbul, Türkiye*

Abstract

Generative models are routinely compared through the distance between a real and a synthetic sample, yet the sample size needed to resolve a difference of interest is rarely determined. We develop a framework for designing and reporting evaluations based on the plug-in 2-Wasserstein distance. Its central object is the resolution floor, the distance between two disjoint real samples of equal size, which is strictly positive and decays at a rate set by the intrinsic dimension of the data. Inverting the fitted decay gives a prospective sample-size rule. Because the rule is valid only where a single power law describes the decay, we report the size, power and severity of the test for that condition. The floor also yields a floor-corrected skill score, a two-sided test that exposes memorisation, a minimum detectable difference at stated error rates and an equivalence test. For latent-generative Wasserstein certificates, the floor determines the dimension above which no certificate of the standard form can be informative; we tabulate this limit for up to 32 768 samples and replace the decoder's Lipschitz constant by the transport cost of the latent optimal coupling, which on trained decoders is within $1.15\times$ of the exact distance, whereas spectral-norm certificates exceed it by factors of 12–120. On eight tabular benchmark encodings, the smallest resolvable difference at 4 000 samples per side ranges from 1.4% of the attainable skill range to 4.5 times that range, so resolution has to be determined for each benchmark.

Keywords: generative model evaluation, optimal transport, Wasserstein distance, intrinsic dimension, power analysis, sample-size planning, tabular data, reproducibility

1 Introduction

A generative model is usually evaluated by drawing a synthetic sample, computing a discrepancy between it and a sample of real data, and comparing the result with the discrepancies of competing models. The Fréchet Inception Distance (Heusel et al., 2017) plays this role for images; for tabular data, practice combines column-wise statistics, pairwise dependence scores and joint distances such as the Wasserstein distance (Patki et al., 2016; Yang et al., 2024). In every case the reported number is an estimate computed from finite samples, and two of its statistical properties are well documented. Its expectation depends on the sample size in a way that differs between models (Chong and Forsyth, 2020), and its sampling variability can be as large as the differences that are reported between models (Card et al., 2020; Bouthillier et al., 2021).

These properties raise a design question that precedes any comparison: at a given sample size, which differences can the evaluation resolve, and how many samples are needed to resolve a difference of a given size? In experimental design this question is answered by a power analysis carried out before data are collected. Existing remedies in generative evaluation work after the fact, by correcting

a score for its bias or attaching error bars to it; none computes, from the real data alone and before a model is trained, the sample size required to resolve a target difference.

This paper develops such a design step for evaluations based on the plug-in 2-Wasserstein distance W_2 . The starting point is an elementary observation. Two independent samples from the same distribution are at a strictly positive W_2 distance from each other, and the typical size of that distance, which we call the *resolution floor*, decays with the sample size n approximately as $n^{-1/d}$, where d is an intrinsic dimension of the distribution (Weed and Bach, 2019). The floor can be measured on real data alone by repeatedly splitting a data set into two disjoint halves. Its level, its rate of decay and its variability then determine three quantities of practical interest: the sample size at which the floor falls below a target, the smallest difference between two models that an evaluation can resolve, and whether a certificate of the standard latent-generative form can be informative at the available sample size.

A concrete case shows why this step matters. On the numeric columns of the `adult` census benchmark with 250 samples per side, the distance of a generator that samples every column independently of the others differs from the floor by 0.017, while its standard deviation across random seeds is 0.222. At this sample size an evaluation cannot distinguish a generator that removes every dependence between the columns from a generator that reproduces the data distribution exactly.

A sample-size rule obtained by extrapolating a fitted power law is only as reliable as the assumption that a single power law describes the decay. We therefore treat the test of this assumption as a statistical instrument in its own right, and we measure its size, its power and the severity of its non-rejections through the same implementation that practitioners use.

1.1 Contributions

- (i) **A design calculus for generative evaluation** (Sections 3 and 4). We define the resolution floor (Definition 1), show that it is non-degenerate for W_2 and degenerate for unbiased statistics such as the maximum mean discrepancy (Proposition 1), and invert the fitted floor law into a sample-size prescription whose sensitivity to the fitted parameters we derive (Proposition 2). The prescription is reported as an interval with measured coverage, and no prescription is returned when the decay exponent is not identified.
- (ii) **Operating characteristics of the specification test** (Sections 4.2 to 4.5). We show that goodness of fit cannot detect departures from a single power law, adopt a parametric-bootstrap (Efron, 1979) curvature test, and report its size as a function of the rate regime, its power as a function of the precision of the floor measurement, and the severity of each non-rejection.
- (iii) **Floor-corrected reporting** (Section 5). We define a skill score relative to a hierarchy of reference generators, show that a one-sided floor test rewards memorisation and replace it by a two-sided test with a null-calibrated leakage screen, and define the minimum detectable difference and an equivalence test at stated error rates.
- (iv) **Limits of latent-generative certificates** (Sections 6 and 7). We identify the statistical term of a standard certificate decomposition as the component that no computation can reduce, give a vacuity criterion (Proposition 3), tabulate the resulting dimension limit for $N \leq 32\,768$ in the units of a measured dimension estimate, test its transfer to non-isotropic, lattice and heavy-tailed laws, and replace the Lipschitz step of the certificate by an exactly computable coupling cost (Lemma 1 and Proposition 4).

- (v) **An empirical study** (Section 8). On eleven table-encoding cells from nine public benchmarks we report fitted floor laws, the measured scope of the sizing rule, a resolution atlas computed without training any model, and certification margins.

The software implementing every procedure, the committed results from which every number in this paper is computed, and machine-checked proofs of the formal statements are described in Section 9.

1.2 Organisation

Section 2 reviews related work. Section 3 introduces the resolution floor and its rate. Section 4 derives the sample-size rule and characterises the test that licenses it. Section 5 turns the floor into reporting tools and summarises the resulting procedure in Figure 2. Sections 6 and 7 treat latent-generative certificates. Section 8 reports the benchmark study, Section 9 the implementation, and Section 10 discusses scope and limitations. Proofs are collected in Appendix A and experimental details in Appendix B.

2 Related Work

Bias and variability of generative metrics. Chong and Forsyth (2020) show that the Fréchet Inception Distance and the Inception Score (Salimans et al., 2016) are biased in finite samples and that the bias depends on the model, so that one model can appear better than another because its bias is smaller; they correct for it by extrapolating to infinite sample size. Jayasumana et al. (2024) point out that the Fréchet distance assumes Gaussian features, and Chen (2026) shows that a statistic built from the first two moments can be gamed: on ImageNet, images constructed to reproduce nothing but the mean and covariance of the reference features score a Fréchet distance of 24.7, better than the 58.6 of held-out real images. Diagnostic refinements separate fidelity from coverage (Sajjadi et al., 2018; Kynkäänniemi et al., 2019). Dufour et al. (2026) measure the Fréchet distance across both training and sampling seeds: independently retrained networks spread $3.2\times$ more widely than repeated draws from one network. These works quantify the bias and the noise of a metric once models exist; the present paper uses the first of the two, the finite-sample bias of the discrepancy, to plan an evaluation before any model is trained.

Evaluation of tabular generators. Tabular generators range from conditional adversarial networks (Xu et al., 2019) to language models (Borisov et al., 2023), and are evaluated by combinations of per-column and pairwise statistics (Yang et al., 2024), models trained on synthetic data (Patki et al., 2016; Esteban et al., 2017) and privacy proxies such as the distance from each synthetic record to its closest real record (Park et al., 2018). The closest prior work is Yang et al. (2024), who decompose fidelity into marginal, pairwise, leave-one-out and joint components and compare every score with three model-free references, including one computed from two halves of the real data. Their framework omits distance-based statistics and does not address sample size. We adopt their references (Section 5.1) and add the measurement theory required for design.

Power and design in machine-learning evaluation. According to Card et al. (2020), a machine-translation test set of the common size of 2 000 sentences detects a one-point BLEU difference with a probability of only about 75%. Bouthillier et al. (2021) find that the sampling of the data, the initialisation of the parameters and the choice of hyperparameters each change benchmark results

substantially, and Miller (2024) provide standard errors, paired tests and power calculations for language-model evaluations; we follow the latter in stating every detectability claim at an explicit significance level and power. Morris et al. (2019) give the framework for simulation studies that we follow when reporting Monte-Carlo error. Classifier two-sample tests (Friedman, 2003; Lopez-Paz and Oquab, 2017) replace the discrepancy by a classifier’s accuracy and face the same sample-size question.

Rates and limit laws for empirical optimal transport. Dudley (1969) established the $n^{-1/d}$ rate for the Wasserstein distance between a measure and its empirical version, Ajtai et al. (1984) the $\sqrt{\log n/n}$ rate in two dimensions, and Fournier and Guillin (2015) non-asymptotic bounds. Weed and Bach (2019) express the rate through an upper Wasserstein dimension, which can be much smaller than the ambient dimension; Niles-Weed and Rigollet (2022) study the resulting limits of estimation in high dimension, and Hundrieser et al. (2024) show that the empirical transport cost between two different measures converges at the rate set by the less complex of the two. Limit laws for the empirical transport cost are available in one dimension (del Barrio et al., 1999), in general dimension under moment conditions (del Barrio and Loubes, 2019), on finite and countable spaces (Sommerfeld and Munk, 2018; Tameling et al., 2019) and for the entropy-regularised problem (Bigot et al., 2019); Sommerfeld and Munk (2018) also show that the naive bootstrap fails for these statistics.

Two-sample tests and model comparison. The maximum mean discrepancy (Gretton et al., 2007, 2012) and its adaptive variants (Schrab et al., 2023) are unbiased two-sample statistics with a classical power theory. Model-versus-model comparison against a shared reference sample is addressed by the relative-similarity test of Bounliphone et al. (2016) and a relative Kullback–Leibler version (Gao et al., 2025). The tests of this paper are screening and design devices; formal comparisons of two models should use those procedures.

Certificates for latent-generative models. Dao et al. (2023) and Lee et al. (2025) bound the Wasserstein distance between the data distribution and the distribution of a latent-generative model by a sum of a statistical term, a latent transport term scaled by the decoder’s Lipschitz constant, and a sampling term. Exact computation of a network’s Lipschitz constant is NP-hard (Virmaux and Scaman, 2018); practical certificates use products of layer spectral norms (Szegedy et al., 2014) or semidefinite relaxations (Fazlyab et al., 2019). A bound is informative only if it is smaller than a trivial bound (Dziugaite and Roy, 2017); we apply this standard in Section 6.

3 The Resolution Floor

This section defines the floor, establishes when it is non-degenerate, relates its decay to intrinsic dimension and describes how it is computed.

3.1 Notation

Let P be a Borel probability measure on $\mathbb{R}^p$ with finite second moment, and let $X_n = \{x_1, \dots, x_n\}$ be an independent sample from P with empirical measure $\mu_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$. For probability measures μ and ν on $\mathbb{R}^p$ with finite second moments, the 2-Wasserstein distance, the quadratic-cost case of the transport problem formulated by Kantorovich (1942), is

$$W_2(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} \int \|u - v\|^2 d\pi(u, v) \right)^{1/2},$$

where $\Pi(\mu, \nu)$ is the set of couplings of μ and ν , that is, of probability measures on $\mathbb{R}^P \times \mathbb{R}^P$ with marginals μ and ν . For two samples X_n and X'_n of equal size we write $D(X_n, X'_n) = W_2(\mu_n, \mu'_n)$ for the plug-in two-sample statistic. For a measurable map g , $g\#\mu$ denotes the pushforward of μ under g . Throughout, N is the number of available real observations, n is the number of observations per side of a comparison, with $2n \leq N$, and n_{rep} is a number of independent replicates. The scale of the data is $\tau = (\text{tr Cov } P)^{1/2}$, the root-mean-square distance of an observation from the mean.

3.2 Definition

Definition 1 (Resolution floor). Given N real observations and a per-side size n with $2n \leq N$, draw two disjoint subsamples of size n uniformly at random, compute D between them, and repeat independently n_{rep} times. The *resolution floor* $f(n)$ is the median of the n_{rep} replicate values, and the *replicate distribution* is their empirical distribution.

The two subsamples must be disjoint: shared observations pull D towards zero and understate the floor. The replicate distribution approximates the null distribution of D when both samples come from the same law (Ramdas et al., 2017). A generated sample whose distance to real data lies inside this distribution is, at size n , statistically indistinguishable from a second real sample; Section 5 develops the resulting tests. Monte-Carlo p -values computed from n_{rep} replicates take the form $(k + 1)/(n_{\text{rep}} + 1)$ (Hope, 1968; Davison and Hinkley, 1997), so the smallest attainable p -value is $1/(n_{\text{rep}} + 1)$ and at least $n_{\text{rep}} = 19$ replicates are needed before a p -value of 0.05 can be reached.

3.3 Existence of the Floor

The design procedures of Section 4 use the rate at which the floor decays, so they require the floor to be strictly positive. The following proposition shows that this holds for W_2 and fails for unbiased discrepancy estimators.

Assumption A1 (Sampling). X_n and X'_n are independent samples of equal size n drawn from a common law P on $\mathbb{R}^P$ with $\int \|x\|^2 dP(x) < \infty$.

Assumption A2 (Non-degeneracy). P has no atoms.

Proposition 1 (Existence of the floor). *Under Assumptions A1 and A2, the plug-in two-sample statistic satisfies*

$$\mathbb{E}[D(X_n, X'_n)] > 0 = W_2(P, P) \quad \text{for every } n \geq 1.$$

If instead $\hat{D}$ is an unbiased estimator of a discrepancy that vanishes when both laws are equal, so that $\mathbb{E}[\hat{D}(X_n, X'_n)] = 0$ under Assumption A1, and $\hat{D}$ is not almost surely constant, then $\hat{D} < 0$ with positive probability, and $\log \max(\hat{D}, 0)$ is undefined on an event of positive probability.

The positive expectation in the first statement is the finite-sample bias of the plug-in estimator, and by Weed and Bach (2019) it decays at a rate governed by an intrinsic dimension of P (Section 3.4). The bias is therefore the object of interest here: its rate carries the dimensional information that the design procedures use.

The second statement explains why the framework is built on W_2 . Unbiased U -statistics such as the energy distance (Baringhaus and Franz, 2004; Székely and Rizzo, 2013) and the maximum mean discrepancy (Gretton et al., 2007, 2012; Schrab et al., 2023) have expectation zero under the null, and a large fraction of their null replicates is non-positive: on a standardised eight-dimensional Gaussian

copula with $N = 4000$ (Appendix B.1), between 62% and 70% of the energy-distance replicates and between 60% and 69% of the MMD replicates are clipped at zero at $n \in \{125, 500, 2000\}$, so the logarithm of their floor is undefined. For such statistics the relevant design quantity is the null variance, and the classical power analysis of Gretton et al. (2012) applies. Sliced W_2 (Rabin et al., 2012; Bonneel et al., 2015) is biased but its bias does not decay at the dimension-governed rate: on the six-dimensional copula of Section 3.4 it gives a decay exponent of 2.07, where exact W_2 on the same data gives 4.50.

3.4 Rate of Decay and the Effective Exponent

For the 2-Wasserstein distance, Weed and Bach (2019) define an upper Wasserstein dimension $d_2^* \geq 4$ and show that the expected one-sample distance satisfies $\mathbb{E} W_2(\mu_n, P) \lesssim n^{-1/s}$ for every $s > d_2^*$. Faster decay is possible: in one dimension, under conditions on the density and its tails, the distance decays at the parametric rate $n^{-1/2}$ (Bobkov and Ledoux, 2019). We fit a power law in n to measured floors.

Definition 2 (Floor law and effective exponent). Let $\mathcal{G} = \{n_1 < \dots < n_k\}$ be a geometric grid of per-side sizes with $2n_k \leq N$. The *effective exponent* d_{eff} and the constant C are the weighted least-squares solution of

$$\log f(n) = \log C - \frac{1}{d_{\text{eff}}} \log n, \quad n \in \mathcal{G}, \quad (1)$$

with weights $f(n)^2/s(n)^2$, where $s(n)$ is the standard deviation of the replicate distribution at n . The fitted relation $f(n) = Cn^{-1/d_{\text{eff}}}$ is the *floor law*. Unless stated otherwise, $\mathcal{G}$ consists of the six sizes $N/64, N/32, \dots, N/2$.

Least squares on logarithmic axes is appropriate here because f is a function of n observed with noise; the pitfalls of fitting power-law probability distributions in this way (Clauset et al., 2009) do not arise. As an illustration, on a six-dimensional Gaussian copula with $N = 4000$ the fitted floor law is $f(n) = 3.7096 n^{-1/4.50}$ with coefficient of determination $R^2 = 0.997$; the exponent is below the ambient dimension 6, as expected for a correlated law. Three properties delimit the interpretation of d_{eff} .

- (a) *The effective exponent describes a finite range of sample sizes.* Where a single exponent governs the decay over $\mathcal{G}$, d_{eff} estimates the dimension in the Weed–Bach rate. Measures that look different at different scales have no single exponent (Weed and Bach, 2019), and for them no scalar exponent can be extrapolated. Section 4.2 gives a test for this case.
- (b) *Low dimensions saturate at an apparent exponent of 2.* The parametric rate $n^{-1/2}$ of one-dimensional laws (Bobkov and Ledoux, 2019) corresponds to an exponent of 2. Measured on isotropic Gaussians with $N = 4096$, the slope estimator of Definition 3 returns 2.12 ± 0.51 (mean and standard deviation over six seeds) for nominal dimension 1, and 2.63, 3.22, 4.19, 6.06, 7.70 and 10.78 for nominal dimensions 2, 3, 4, 6, 8 and 12: relative biases of +112%, +32%, +7%, +5%, +1%, −4% and −10%. Upward bias in low dimension reflects the saturation; the downward bias in high dimension reflects the finite range of the fitting window.
- (c) *Three rate regimes.* The Weed–Bach dimension d_2^* is defined as an infimum over exponents larger than 4, so a fitted exponent below 4 is not an estimate of it. We call $d_{\text{eff}} < 2$ the parametric

regime, $2 \leq d_{\text{eff}} < 4$ the crossover regime, in which the decay need not follow a pure power law, and $d_{\text{eff}} \geq 4$ the dimension-governed regime. The sample-size rule of Section 4 is used only in the dimension-governed regime.

The same decay yields a direct estimate of dimension from a single data set, which Section 6 uses.

Definition 3 (Slope dimension estimate). For a sample of size N , the *slope dimension estimate* $\hat{d}$ is $-1/b$, where b is the least-squares slope of $\log D$ against $\log n$ over three random disjoint splits at each of $n \in \{N/8, N/4, N/2\}$.

Estimating intrinsic dimension from the decay of the Wasserstein distance between independent empirical samples was proposed by Block et al. (2022), who divide $\log \alpha$ by the drop in $\log W_1$ between sample sizes n and αn and prove non-asymptotic guarantees for this estimator. Definition 3 applies their construction to W_2 , with a least-squares slope over three sample sizes.

3.5 Computation by Exact Optimal Transport

For two samples of equal size with uniform weights, the Kantorovich linear program (Kantorovich, 1942) has an optimal solution that is a permutation: by the Birkhoff–von Neumann theorem (Birkhoff, 1946; von Neumann, 1953), every doubly stochastic matrix is a convex combination of permutation matrices, and the transport cost is linear in the plan. The distance therefore reduces to a linear assignment problem, which shortest-augmenting-path methods descending from the Hungarian method of Kuhn (1955) solve exactly in $O(n^3)$ operations (Jonker and Volgenant, 1987); we use the implementation of Crouse (2016), and Peyré and Cuturi (2019) survey the alternatives. We use exact assignment throughout. Entropic regularisation (Cuturi, 2013) adds a bias that depends on the regularisation strength and would be confounded with the finite-sample bias being measured. Between two samples of a 64-dimensional standard normal law, sliced W_2 underestimates W_2 by 99%, and the sharp comparison between the sliced and the standard 1-Wasserstein distance of compactly supported measures has the exponent $1/d$ (Carlier et al., 2025), which makes it uninformative in high dimension; linear comparisons need strong structural assumptions, for example a discrete measure whose optimal coupling moves every point to a nearest atom (Niles-Weed and Shkrob, 2026).

Exact transport is affordable at the sizes used in this paper. One exact solve on one processor core takes about 2.1 seconds at $n = 4\,096$ and 12 seconds at $n = 8\,192$, with an empirical growth exponent in n between 2.5 and 2.6 (Appendix B.2). For N real observations, the largest size on the fitting grid is $n = N/2$.

4 Designing an Evaluation

This section derives a sample-size rule from the floor law, states the condition under which the rule is valid, and characterises the test of that condition by its size, its power and the severity of its non-rejections.

4.1 Two Inversions of the Floor Law

The floor law of Definition 2 can be inverted in two ways. The first gives the per-side size at which the floor reaches a target level ϕ ,

$$n^*(\phi) = (C/\phi)^{d_{\text{eff}}}. \quad (2)$$

The second gives the smallest change in expected distance that two independent evaluations at size n can resolve at significance level α and power $1 - \beta$,

$$\Delta(n) = (z_{1-\alpha/2} + z_{1-\beta}) \sqrt{2} s(n), \quad (3)$$

where $s(n)$ is the standard deviation of the replicate distribution and z_q the standard normal quantile. Equation (3) requires the variability of the floor to be small relative to the effect of interest; it does not require the floor itself to be small. For the six-dimensional Gaussian copula of Section 3.4, with $s(n)$ extrapolated along the floor law at the median coefficient of variation of the fitting grid, the resolvable change at $\alpha = 0.05$ and power 0.80 is 0.144 at $n = 500$ and 0.086 at $n = 5000$; a change of 0.043, about a third of the first, requires $n \approx 114\,000$. Section 5.4 extends Equation (3) to differences between models.

Equation (2) inverts fitted parameters, and its sensitivity to them is exact.

Proposition 2 (Sensitivity of the prescription). *Fix $\phi < C$ and let n^* be given by Equation (2). Then*

$$\frac{\partial \log n^*}{\partial \log C} = d_{\text{eff}}, \quad \frac{\partial \log n^*}{\partial d_{\text{eff}}} = \log(C/\phi) = \frac{\log n^*}{d_{\text{eff}}}.$$

Consequently a relative perturbation ε of the constant multiplies the prescription by $(1 + \varepsilon)^{d_{\text{eff}}} \approx e^{d_{\text{eff}} \varepsilon}$, and a relative perturbation δ of the exponent multiplies it by $(n^*)^\delta$.

Proposition 2 has three consequences for design. First, the elasticity of the prescription with respect to the exponent equals $\log n^*$, so the same relative error in d_{eff} costs more the further the target lies beyond the fitted grid. Second, a point prescription conveys false precision, and Section 4.6 replaces it by an interval. Third, the exponent must be identified before it is inverted: where no single exponent governs the decay, d_{eff} is not a parameter of the data and Equation (2) has no meaning.

Assumption A3 (Single-exponent regime). Over the grid $\mathcal{G}$, a single exponent governs the decay of f , so that Equation (1) holds up to replicate noise, and $d_{\text{eff}} \geq 4$.

4.2 Testing for a Single Exponent

The coefficient of determination of the fit is not a suitable test of Assumption A3. Consider the floor $f(n) = n^{-1/2} + 0.35 n^{-1/8}$, whose local exponent changes continuously with n . With realistic replicate noise, the probability that $R^2 < 0.95$ is 0.14 on a four-point grid, 0.007 on six points and 0.000 on eight: adding points to a smooth curve improves the fit and makes the departure harder to detect. A fixed threshold on the ratio of the exponents fitted to the two halves of the grid fails in the opposite direction: on a four-point grid, a threshold of $1.25\times$ rejects 62% of exact power laws at a 3% floor noise level. We therefore calibrate the drift statistic against a simulated null.

Definition 4 (Curvature test). Let d_{low} and d_{high} be the effective exponents fitted by the weighted least squares of Definition 2 to the lower and upper halves of $\mathcal{G}$, and let the *drift* $\hat{\delta} = \max(d_{\text{low}}, d_{\text{high}}) / \min(d_{\text{low}}, d_{\text{high}})$. Draw B synthetic floor profiles from the fitted law $Cn^{-1/d_{\text{eff}}}$ with independent lognormal perturbations whose standard deviation at each grid point equals the bootstrap standard error of $\log f(n)$ over the replicates, and compute their drifts $\hat{\delta}^{(1)}, \dots, \hat{\delta}^{(B)}$. The curvature p -value is $(1 + \#\{b : \hat{\delta}^{(b)} \geq \hat{\delta}\}) / (B + 1)$. A floor profile is *admissible* when $R^2 > 0.95$ and the curvature p -value is at least 0.05.

The bootstrap standard error of the median is used because the floor is a median; the standard error of a mean understates the variability of a median and inflates the rejection rate. When a profile is not admissible, no prescription is returned. In its place we report the range of prescriptions obtained from the low-half, high-half and full fits, which shows how strongly the answer depends on the part of the grid that is extrapolated. The same holds for $d_{\text{eff}} < 4$, where Assumption A3 fails by definition.

4.3 Size of the Test

We measured the rejection rate of the curvature test on isotropic Gaussian laws, for which a single exponent holds exactly, with $N = 1\,024$ and $n_{\text{rep}} = 12$ (Table 1, Figure 1a).

nominal dimension	2	3	4	6	8
d_{eff} of the null law	2.69	3.43	4.27	5.92	7.44
rejection rate at nominal 0.05	0.105	0.120	0.025	0.040	0.023
mean $\log(d_{\text{low}}/d_{\text{high}})$	+0.120	+0.081	+0.047	−0.007	−0.037

Table 1: Size of the curvature test against exact single-exponent laws (isotropic Gaussians, $N = 1\,024$, $n_{\text{rep}} = 12$; 200 trials at nominal dimensions 2–4 and 175 at 6 and 8), with the mean signed log drift of the null profiles.

The rejection rate exceeds the nominal level in the crossover regime and lies below it in the dimension-governed regime. The mean signed drift explains the pattern. It is systematically positive for $d_{\text{eff}} < 4$ and changes sign between $d_{\text{eff}} = 4.27$ and 5.92 , falling from $+0.120$ at $d_{\text{eff}} = 2.69$ to -0.007 at 5.92 . Below an exponent of 4 the rate theory gives no pure power law for W_2 : the bounds of Fournier and Guillin (2015) change form at dimension 4, and in two dimensions the decay carries a logarithmic factor (Ajtai et al., 1984). A single power law is therefore a misspecified null there, and the excess rejections are detections of a real departure from it. In the regime where Assumption A3 can hold, the test is conservative.

4.4 Power and Floor Precision

We measured the power of the test against the two-scale floor $n^{-1/2} + 0.35 n^{-1/8}$ at three levels of floor precision, expressed as the median standard error of $\log f(n)$ over the grid. These levels cover most of the precision achieved on the eight admissible benchmark cells of Section 8, whose median standard errors lie between 0.3% and 5.3% at $n_{\text{rep}} = 12$. Table 2 and Figure 1b report power for grids of the same range with more points and for grids of wider range.

Power is governed by the precision of the floor measurement and by the range of the grid. Adding points within a fixed range changes it little, whereas extending the range raises it substantially. The standard error of the median floor scales approximately as $\text{SE} \approx 1.25 \cdot \text{CV} / \sqrt{n_{\text{rep}}}$, where CV is the replicate coefficient of variation, so precision can be bought with replicates. The range is limited by the number of real observations through $2n \leq N$; on a table of fixed size, power can therefore be increased only by adding replicates, and a larger table buys range.

median SE of $\log f$	range [125, 4 000]		wider range	
	6 points	12 points	8 points, to 16 000	12 points, to 256 000
1%	0.885	0.975	1.00	1.00
3%	0.18	0.26	0.335	0.965
5%	0.08	0.105	0.165	0.635

Table 2: Power of the curvature test at nominal size 0.05 against the two-scale floor $n^{-1/2} + 0.35 n^{-1/8}$ ($n_{\text{rep}} = 12$, 200 trials per cell).

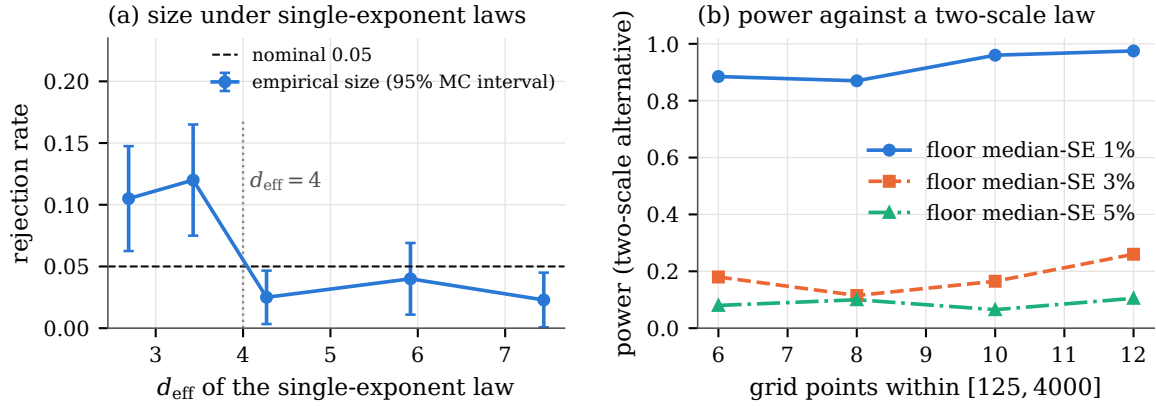


Figure 1: The curvature test as a measuring instrument. (a) Rejection rate against exact single-exponent laws with 95% Monte-Carlo intervals: the rate exceeds the nominal level below $d_{\text{eff}} = 4$, where a single power law is misspecified. (b) Power against a two-scale alternative as a function of the number of grid points within [125, 4 000], at three floor precisions.

4.5 Severity of a Non-Rejection

A non-rejection by a test of limited power does not establish that the decay follows a single power law. What it does establish depends on the precision of the floor, and we report it as a severity threshold in the sense of the smallest departure the test would have detected.

Definition 5 (Severity threshold). For an admissible floor profile with fitted (C, d_{eff}) and a drift factor $\delta \geq 1$, consider the alternative floor with exponents $d_{\text{eff}}\sqrt{\delta}$ on the lower half of $\mathcal{G}$ and $d_{\text{eff}}/\sqrt{\delta}$ on the upper half, joined at the midpoint of the grid and perturbed by the measured noise of the profile. The *severity threshold* δ_{50} is the drift at which the rejection probability of the curvature test equals 1/2, obtained by linear interpolation of Monte-Carlo rejection rates in $\log \delta$.

Drifts below δ_{50} remain undetected at least half of the time, so a non-rejection is informative only about departures larger than δ_{50} . The Monte-Carlo estimate of δ_{50} has two sources of error: the binomial error of the rejection rates within one simulation seed, and the variation of the whole procedure between seeds. We estimate both from eight seeds and report a combined 90% interval. On seven of the eight admissible benchmark cells of Section 8 the between-seed component is the

larger of the two, and the combined interval is $1.38\text{--}2.19\times$ wider than the binomial interval alone, with a median of $1.65\times$.

4.6 The Prescription as an Interval

Following Proposition 2, we propagate the estimation error of the fitted pair $(\log C, 1/d_{\text{eff}})$ through Equation (2). We draw from a multivariate t distribution centred at the weighted least-squares estimate with its estimated covariance and report quantiles of the resulting prescriptions. The coverage of this interval was measured on exact power laws with known parameters: at nominal level 0.90 it delivers between 0.860 and 0.920, with a median of 0.880 slightly below the nominal level. A calibrated version requests the nominal level 0.9125, which delivers 0.90 as a median over the regime in which it was measured: six-point grids over $[125, 4\,000]$, $n_{\text{rep}} = 12$, $d \in \{4, 5, 6\}$ and replicate coefficients of variation in $\{0.03, 0.08, 0.13\}$. Away from this regime the calibration is not applied, and the uncalibrated interval is returned with a flag; the regime in which the software applies it is stated with every interval.

The coverage refers to independent replicate noise at each grid size. On real data all grid sizes are subsampled from one table with shared random numbers, which reduces the variance of the drift statistic to 0.81 times its value under independent replicates; the construction is therefore, if anything, conservative on real data. The interval is conditional on the data set at hand and does not describe how much a second data set from the same law would change the answer. That variation is not negligible: on the benchmark tables of Section 8, the slope dimension estimate varies 1.1 to 2.1 times more across half-size subsamples of a table than across re-splits of the whole table, with a median of 1.5.

5 Reporting an Evaluation

A designed evaluation must also report its outcome on a scale that is comparable across sample sizes and robust to degenerate generators. This section defines such a scale and the associated tests.

5.1 Reference Generators and the Skill Score

A raw distance has no meaning on its own: a value of 0.1 indicates an excellent generator if the floor at that sample size is 0.09 and a poor one if it is 0.001. We therefore compare every generator with reference generators that are computed from the real data and require no training.

- **T0** places all mass at the real mean. Its distance to the real sample is the data scale τ , and it is the best constant generator.
- **T1** resamples each column independently. It reproduces all marginal distributions and no dependence between columns; it coincides with the perm reference of Yang et al. (2024).
- **T2** is the Gaussian law with the real mean and covariance. It reproduces first and second moments and nothing beyond them.
- **T_{self}** resamples real observations that are not in the reference subsample, as the self reference of Yang et al. (2024). It approximates an ideal generator and serves as a ceiling. Drawn from the disjoint remainder, its distance is $1.07\times$ the floor; resampled from the whole data set, it falls below the floor, to $0.88\times$, because it shares observations with the reference (eight-dimensional Gaussian copula, $n = 500$, median over 20 seeds).

Definition 6 (Skill score). Let D_{obs} be the distance between a generated sample and a real reference sample of the same size n , let $f = f(n)$, and let D_B be the distance of a reference generator B . The *floor-corrected ratio* is

$$V = \frac{D_{\text{obs}} - f}{D_B - f}, \quad (4)$$

and the *skill* relative to B is $1 - V$.

The skill equals 1 for a generator that is as close to the real sample as a second real sample, and 0 for a generator as distant as the reference B . Subtracting the floor from numerator and denominator makes V dimensionless. For a fixed model law $Q \neq P$, $V \rightarrow W_2(P, Q)/W_2(P, Q_B)$ as $n \rightarrow \infty$, which does not depend on n , and for $Q = P$, $V \rightarrow 0$; at finite n the subtraction is a first-order correction, and V still varies with n . On an eight-dimensional Gaussian copula with non-Gaussian marginals at $n = 500$ per side, the median skill relative to T1 over 20 seeds is +1.00 for a fresh draw from the true law (range 0.96 to 1.03), +0.83 for a sample of T2, +0.00 for a sample of T1 and −0.52 for a sample of T0.

5.2 A Two-Sided Floor Test

A one-sided test that asks only whether D_{obs} exceeds the floor rewards a generator that returns the training observations. Such a sample shares points with the reference sample, its distance falls below the floor, and the one-sided test finds no evidence against it. On a six-dimensional Gaussian copula, an exact copy of the training data obtains a skill of +1.42 relative to T2, above the nominal maximum of 1, whereas a fresh draw from the true law obtains +0.91. Chen (2026) reports the analogous phenomenon for image metrics.

If the generator draws fresh samples from P , D_{obs} is an ordinary draw from the replicate distribution and should lie in neither of its tails. We therefore report both tail probabilities, $p_{\text{hi}} = \mathbb{P}(D' \geq D_{\text{obs}})$ and $p_{\text{lo}} = \mathbb{P}(D' \leq D_{\text{obs}})$, where D' follows the replicate distribution, as Monte-Carlo p -values. A small p_{hi} indicates a generator distinguishable from real data, and a small p_{lo} indicates a sample that is closer to the reference than a second real sample would be, the signature of leakage. A skill above 1 is not itself evidence of leakage: for a generator that draws from P , the probability of a skill above 1 is close to one half, 0.52 at $d = 4$ and 0.42 at $d = 8$ in 200 trials each, and a rule flagging $V < -0.05$ against T2 would flag 24–28% of such generators. The lower-tail probability is the calibrated statistic.

The skill score also needs its own interval, because its denominator $D_B - f$ is itself random. On the numeric columns of `adult` at $n = 250$, the standard deviation of D_{T1} across random seeds is 0.222, which is 13× the difference $D_{\text{T1}} - f = 0.017$ that forms the denominator.

5.3 Null-Calibrated Leakage Screening

A memorising generator produces a skill value that says nothing about fidelity, so leakage screening is part of the evaluation. We split the real sample into disjoint halves A and B . Because B is a genuine sample from P that shares no observation with A , the distribution of distances from points of B to their nearest neighbours in A , and the fraction of exact duplicates, describe what a generator drawing fresh samples from P produces. The generated sample is compared with this null, so the screen needs no threshold chosen in advance. It detects an exact memoriser of a table with 60 000 rows and raises no alarm on any of four same-law generators for discrete laws, where real observations legitimately coincide.

Two limits apply to every screen of this kind. A memoriser that perturbs each copied record by more than about $0.44\times$ the median distance between records escapes detection. And when a law has a finite support that the sample covers densely, so that $n/|\text{support}| \approx 1$, memorisation cannot be distinguished from fresh sampling by any detector, including density-based membership-inference attacks that target local overfitting (van Breugel et al., 2023); the remedy is a training split disjoint from the evaluation data.

5.4 The Minimum Detectable Difference

Let two generators be evaluated against the same real reference sample X_r , with distances $D_A = D(X_r, Y_A)$ and $D_B = D(X_r, Y_B)$. Sharing X_r removes the variability of the reference from the difference $D_A - D_B$.

Definition 7 (Minimum detectable difference). At significance level α and power $1 - \beta$, the *minimum detectable difference* at size n is

$$\text{MDD} = (z_{1-\alpha/2} + z_{1-\beta}) \sigma_{\text{pair}}, \quad \sigma_{\text{pair}} = \text{sd}(D(X_r, Y_A) - D(X_r, Y_B)), \quad (5)$$

where the standard deviation is taken over independent draws of the reference sample and of both generated samples of size n .

Every detectability statement in this paper specifies both α and the power. The common practice of calling a difference significant when it exceeds two standard deviations corresponds to power $1/2$: in simulation, a true difference equal to that threshold is detected with frequency 0.51 over 400 trials, against 0.500 analytically.

The paired standard deviation depends on the level of the distances being compared, because the spread of D grows with its mean. Before any model exists it can be estimated in two ways. The *direct estimate* draws repeated real references and pairs of samples from a reference generator at the level of interest, typically T1, and computes their standard deviation. The *floor proxy* scales the replicate variability of the floor to that level, $\sqrt{2} \text{CV}(f) \cdot D_{\text{level}}$, where $\text{CV}(f)$ is the coefficient of variation of the replicate distribution. The proxy requires no additional computation and overstates the paired standard deviation: on a synthetic copula by factors between 1.39 and 1.75 (200 replicates for each of three generators), and on the benchmark tables of Section 8.4 by factors between 1.05 and 10.53. Using the unscaled floor variability instead is not safe: its ratio to the paired spread falls from 1.38 near the floor to 0.79 at the level of T1, so it understates the spread where models are compared. Differences between models are measured on the scale of D , which does not translate linearly into differences between laws: shifting a law by a fixed vector that raises the expected distance by one equivalence margin (Section 5.5) changes the optimal matching and multiplies the standard deviation of D by 2.9.

5.5 Equivalence Testing

A statement that a generator is indistinguishable from real data is an equivalence claim, and it requires an equivalence test. We apply the two one-sided tests procedure (Westlake, 1981; Schuirmann, 1987) to the difference between D_{obs} and the floor, with a margin η fixed before the evaluation, and report η_{min} , the smallest margin that the precision of the floor can support. On a six-dimensional copula, with the margin $\eta = 0.2193$ set to a quarter of the floor and $\alpha = 0.05$, the procedure declares equivalence

Before any model is trained (real data X with N rows) 1. Fix α, the power $1 - \beta$, the equivalence margin η and at least two embeddings of categorical columns (Section 10.1); standardise the columns on X. 2. Measure the floor on the grid $N/64, \dots, N/2$ with $n_{\text{rep}} \geq 19$ disjoint splits, fit the floor law and run the curvature test (Definitions 2 and 4). 3. If the profile is admissible and $d_{\text{eff}} \geq 4$, report the prescription interval (Section 4.6) and the severity threshold (Definition 5); otherwise report the range of the half-grid prescriptions. 4. Report the minimum detectable difference at the planned size, measured at the level of a reference generator or bounded by the floor proxy (Definition 7). 5. For a latent-generative certificate, compare the slope estimate $\hat{d}$ and its tail index with the limit at N (Sections 6.4 and 6.5). After the models are trained 6. For each generated sample, report the distance, the skill against T0, T1 and T2 with its interval, both tail probabilities and the leakage screen (Sections 5.1 to 5.3). 7. Claim indistinguishability only through the equivalence test at the margin η (Section 5.5), and state α and the power with every detectability statement.

Figure 2: Designing and reporting an evaluation based on W_2 .

at the boundaries $\pm\eta$ with frequency 0.0, so its size is at most α ; it declares equivalence for fresh draws from the true law with frequency 1.0; and it rejects the memoriser, whose distance to the reference is zero. For $\eta = 0$ it reduces to the two-sided floor test of Section 5.2. The upper-tail p -value is calibrated for fresh draws: over 200 fresh generators, $\mathbb{P}(p_{\text{hi}} \leq 0.05) = 0.045$ and the median is 0.524. An evaluation whose $\eta_{\min}$ exceeds every margin of interest cannot establish equivalence for any generator; this is a property of its design.

5.6 The Procedure in Brief

Figure 2 collects the steps of Sections 3 to 5 into one procedure. The design steps use the real data only; the significance level, the power, the equivalence margin and the embedding of categorical columns are fixed before any generator is evaluated.

6 Limits of Latent-Generative Certificates

The preceding sections concern what an evaluation can measure. This section concerns what it can prove about a latent-generative model, and shows that the resolution floor limits both.

6.1 Setting

Let Υ be the data distribution on $\mathbb{R}^m$ and Υ_N the empirical measure of N observations from it. A latent-generative model consists of a latent law Σ and a measurable decoder g ; its output law is $\tilde{\Upsilon} = g_{\#}\Sigma$, and $\tilde{\Upsilon}_M$ is the empirical measure of M decoded latent draws. The certificates of Dao et al. (2023) and Lee et al. (2025) rest on the triangle inequality

$$W_2(\Upsilon, \tilde{\Upsilon}) \leq \underbrace{W_2(\Upsilon, \Upsilon_N)}_{\text{statistical term}} + \underbrace{W_2(\Upsilon_N, \tilde{\Upsilon}_M)}_{\text{empirical term}} + \underbrace{W_2(\tilde{\Upsilon}_M, \tilde{\Upsilon})}_{\text{sampling term}}. \quad (6)$$

The sampling term can be made small by drawing more latent samples. The empirical term involves only finite samples and is usually bounded through the Lipschitz constant of the decoder; Section 7 shows that it can be bounded without that constant. The statistical term depends on the unknown distribution Υ and on the number N of available observations, and no computation reduces it. Following Dziugaite and Roy (2017), we call a bound *vacuous* when it is not smaller than the trivial bound, here the data scale τ of Υ .

6.2 A Vacuity Criterion and a Computable Surrogate

Proposition 3 (Vacuity criterion). *Let $S = W_2(\Upsilon, \Upsilon_N)$ be the statistical term of Equation (6) and τ the data scale. If $S/\tau > 1$, then the right-hand side of Equation (6) exceeds τ and every bound of this form is vacuous.*

The statistical term cannot be computed, because it involves the population. Its expected size is approximated by the distance between two independent samples of size N from the same law, which obeys the same rate with a larger constant, heuristically by a factor $2^{1/d^*}$ that tends to 1 as d^* grows. For a law of intrinsic dimension d we therefore use the dimensionless *floor ratio*

$$r(d, N) = \frac{\mathbb{E} D(X_N, X'_N)}{\tau} \quad (7)$$

as a surrogate for $\mathbb{E} S/\tau$. The triangle inequality gives $D(X_N, X'_N) \leq W_2(\mu_N, P) + W_2(P, \mu'_N)$ and hence only $\mathbb{E} S \geq r\tau/2$. The surrogate is closer than this bound suggests: on isotropic Gaussian laws, the ratio of the two-sample distance to the distance between a sample of size N and a large reference sample lies between 1.007 and 1.16, decreasing with d and N (Appendix B.3). The antecedent of Proposition 3 is the statistical term; a crossing of the surrogate implies a crossing of the term up to this measured factor.

6.3 Reference Tables

We measured $r(d, N)$ on isotropic Gaussian laws, for which the intrinsic and the ambient dimension coincide and $\tau = \sqrt{d}$, with exact transport and 20 replicates per cell (Table 3). The ratio rises with d , exceeds 1 between $d = 16$ and $d = 64$, and decreases slowly with N . The standard error of every cell is at most 0.0031 for $d \geq 4$ and at most 0.013 overall. Near the crossing the sensitivity $d \log d / d \log r$ is between 3.9 and 5.5, so a relative error ε in r shifts the crossing by a relative amount of about 5ε in d ; Section 8.5 propagates this.

6.4 Crossings in the Units of a Measured Dimension

On isotropic laws, $r(d, N)$ crosses 1 at nominal dimension $d \approx 25.8, 34.6, 39.9, 44.9, 49.6, 54.0$ and 58.2 for $N = 256, 1\,024, 2\,048, 4\,096, 8\,192, 16\,384$ and $32\,768$. A practitioner does not know the nominal dimension of real data; the available quantity is a measured estimate such as the slope estimate $\hat{d}$ of Definition 3. This estimate is biased upward in low dimension and downward in high dimension: nominal dimension 64 is measured as 36.4 at $N = 2\,048$ and 41.0 at $N = 32\,768$. Comparing a measured estimate with a threshold stated in nominal units therefore overstates the available headroom: at the crossings below, the nominal limit exceeds the measured one by factors between 1.50 and 1.54. We tabulated the mean of $\hat{d}$ on the same isotropic laws (Table 9 in Appendix B.4) and map the crossing into measured units. The resulting limits are $\hat{d} \approx 17.2$ [16.6, 17.8] at $N = 256, 22.8$

d	1	2	4	8	16	32	64	128
$N = 256$	0.158	0.236	0.435	0.681	0.895	1.052	1.161	1.239
$N = 1\,024$	0.075	0.137	0.315	0.568	0.804	0.986	1.117	1.208
$N = 2\,048$	0.052	0.107	0.268	0.519	0.765	0.957	1.097	1.194
$N = 4\,096$	0.037	0.078	0.227	0.475	0.729	0.931	1.078	1.180
$N = 8\,192$	0.031	0.059	0.191	0.435	0.694	0.905	1.060	1.168
$N = 16\,384$	0.021	0.044	0.162	0.399	0.662	0.880	1.042	1.156
$N = 32\,768$	0.013	0.033	0.137	0.366	0.631	0.857	1.025	1.144

Table 3: The floor ratio $r(d, N)$ of Equation (7) on isotropic Gaussian laws of dimension d with N observations per side (exact transport, 20 replicates per cell).

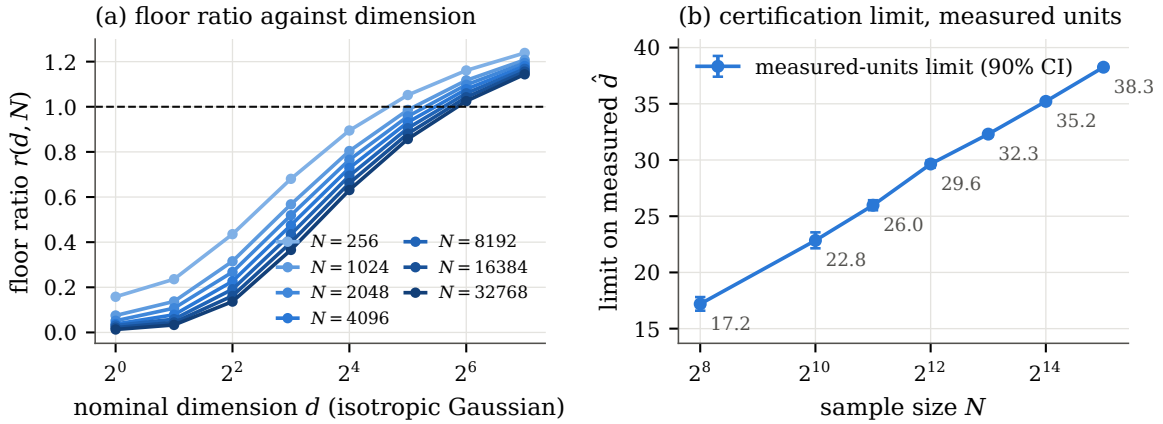


Figure 3: (a) The floor ratio $r(d, N)$ on isotropic Gaussian laws for $N = 256$ to $32\,768$ observations per side; the dashed line is the trivial scale $r = 1$. (b) The resulting certification limit in the units of the slope dimension estimate, with 90% intervals propagating the sampling error of both reference tables.

[22.1, 23.6] at $N = 1\,024$, 26.0 [25.5, 26.4] at $N = 2\,048$, 29.6 [29.3, 30.0] at $N = 4\,096$, 32.3 [32.1, 32.5] at $N = 8\,192$ and 35.2 [35.0, 35.4] at $N = 16\,384$ and 38.3 [38.0, 38.5] at $N = 32\,768$, where the 90% intervals propagate the sampling error of both tables (Figure 3). A data set whose measured dimension exceeds the limit for its size admits no informative certificate of the form Equation (6). The tables cover $N \leq 32\,768$. For larger N the software clamps the limit at its value for $32\,768$, which understates the true limit.

6.5 Transfer to Other Classes of Laws

The reference tables are built on isotropic Gaussian laws, and applying them to real data requires that r depend on the law mainly through its measured dimension. We tested this on three families of non-isotropic laws in 64 ambient dimensions, namely anisotropic Gaussians, factor models and curved manifolds (Appendix B.5), with measured dimensions $\hat{d} \approx 9.7$ – 22.8 , the band in which the crossing lies. Indexed by the measured dimension, the isotropic table predicts the measured floor

ratio with a median residual of -0.4% and a range of -9.9% to $+7.2\%$. The isotropic reference itself, measured with three replicates, varies between -7.5% and $+7.0\%$, so the residuals are within its own measurement noise.

These three families are smooth, unimodal and light-tailed. We tested two further classes, with a verdict rule fixed before the experiment: a class transfers if every residual in the band lies within $\pm 12\%$, the range above widened by the reference noise. Lattice laws transfer: factor models rounded to a grid of step 0.5 or 1 in every coordinate give residuals between -8.7% and $+2.6\%$, and rounding half of the coordinates gives -5.0% and -4.3% . In 64 dimensions these lattices produce no coincident observations, so the test does not cover coarse low-dimensional lattices such as integer-coded categorical columns, where ties reduce the floor (Appendix A, remark after the proof of Proposition 1). Polynomial tails do not transfer: a multivariate Student- t law with $\nu = 3$ degrees of freedom gives a residual of $+44.1\%$ at $N = 1\,024$, and at $N = 256$ its dimension cannot be estimated. With $\nu = 5$ the residuals are $+6.3\%$ in the band and -13.7% outside it, and a rotated Laplace law with exponential tails stays within $\pm 1\%$. A positive residual means that the table understates r , so the certification limit is optimistic for polynomially tailed laws.

We therefore report, with every dimension estimate, the Hill estimate (Hill, 1975) of the tail index of the distances from the median, computed from the largest 5% of those distances, and flag values below 5. On the probe laws the estimate is 2.8 for $\nu = 3$, 4.4 for $\nu = 5$, 9.0 for the Laplace law and 16.1 for an anisotropic Gaussian. The flag is a screening device without a formal guarantee. It is raised on eight of the eleven benchmark cells of Section 8, whose estimates range from 2.75 to 10.94; Section 8.5 accounts for this.

7 A Certificate Without a Lipschitz Constant

The empirical term of Equation (6) compares decoded samples. Published certificates bound it by the latent transport distance multiplied by a Lipschitz constant of the decoder. This section shows that the constant is unnecessary for the empirical statement and that removing it tightens the bound substantially.

7.1 The Plan-Pullback Bound

Lemma 1 (Pushforward coupling). *Let Σ and Σ' be Borel probability measures on a latent space, let $\pi \in \Pi(\Sigma, \Sigma')$, and let g be a measurable map into $\mathbb{R}^m$ with $\int \|g(z)\|^2 d\Sigma(z) < \infty$ and $\int \|g(z)\|^2 d\Sigma'(z) < \infty$. Then $(g \otimes g)_\# \pi$ is a coupling of $g_\# \Sigma$ and $g_\# \Sigma'$, and consequently*

$$W_2(g_\# \Sigma, g_\# \Sigma')^2 \leq \int \|g(z) - g(z')\|^2 d\pi(z, z'). \quad (8)$$

Lemma 1 is elementary (Villani, 2009, Chapter 1); we prove it in Appendix A. The Lipschitz certificate follows from Equation (8) by bounding the integrand by $L^2 \|z - z'\|^2$, where L is a Lipschitz constant of g , and choosing π optimal for the latent cost. For empirical measures this bounding step can be omitted.

Proposition 4 (Empirical plan-pullback bound). *Let σ^* be an optimal assignment, for the squared Euclidean cost, between latent points $z_1, \dots, z_N$ and $z'_1, \dots, z'_N$ with uniform empirical measures*

Σ_N and Σ'_N . Then, for every measurable decoder g ,

$$W_2(g\#\Sigma_N, g\#\Sigma'_N)^2 \leq \frac{1}{N} \sum_{i=1}^N \|g(z_i) - g(z'_{\sigma^*(i)})\|^2, \quad (9)$$

and the right-hand side can be evaluated exactly with $2N$ evaluations of g .

Proposition 4 applies when both samples consist of decoded latent points, as when the observations are represented by their encodings: with Σ_N the empirical measure of the encoded observations and Σ'_N that of N latent draws, the left-hand side of Equation (9) is the empirical term of Equation (6) computed on reconstructions.

The bound holds for every decoder because σ^* defines a valid coupling of the decoded samples, although not necessarily an optimal one. It needs no derivative of g and no Lipschitz constant. The Lipschitz step is also the expensive one: exact computation of a network’s Lipschitz constant is NP-hard (Virmaux and Scaman, 2018), the product of the layer spectral norms (Szegedy et al., 2014) is a loose upper bound, and tighter bounds require semidefinite relaxations (Fazlyab et al., 2019). Optimality of σ^* is not needed for validity; it makes the bound tight when g approximately preserves the latent geometry.

7.2 Tightness on Trained Decoders

For empirical measures of equal size the left-hand side of Equation (9) is itself an assignment problem and can be computed exactly, so the tightness of the bound is a measurable quantity. We trained variational autoencoders (Kingma and Welling, 2014) on the `digits` data set of the UCI repository (Kelly et al., 2026), as distributed with `scikit-learn` (Pedregosa et al., 2011) (1 797 observations of dimension 64), using 1 300 observations for training, in two architectures and with three training seeds each (Appendix B.8). For each trained model and each of two evaluation draws, we compared the encoder means of 400 held-out observations with 400 draws from the latent prior, both decoded by the trained decoder.

Across these twelve cells, the plan-pullback bound is $1.07\text{--}1.15\times$ the exact output-space distance, with a median of $1.09\times$, and never falls below it (Figure 4). The certificate obtained from the product of the decoder’s layer spectral norms is $12\text{--}120\times$ the exact distance, with a median of $53\times$. Replacing the certified constant by the largest Lipschitz ratio observed between sampled latent pairs, which is a lower bound on the true constant and no longer yields a valid certificate, still gives $1.8\text{--}5.1\times$, with a median of $2.9\times$. Most of the looseness of the Lipschitz route therefore comes from replacing the cost of an actual coupling by a worst-case stretch factor, and only part of it from the looseness of the spectral-norm product.

7.3 Scope

Proposition 4 bounds the distance between decoded samples that have been drawn. A Lipschitz constant remains necessary for statements about latent points that have not been drawn, such as a guarantee that holds uniformly over the latent space, and for outputs too expensive to decode. The tightness figures above are measurements on one data set with two architectures; the validity of the bound holds for every decoder.

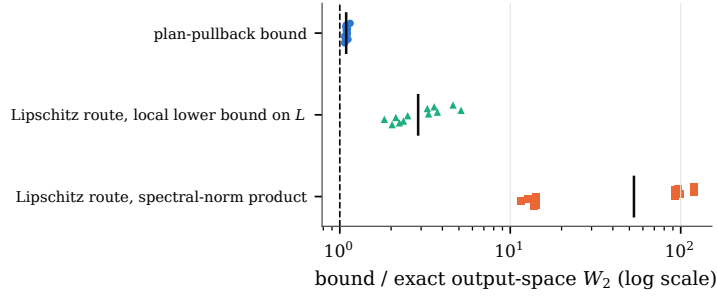


Figure 4: Bounds on the output-space W_2 between decoded held-out encodings and decoded prior draws, divided by the exact distance (log scale). Twelve cells: two architectures, three training seeds and two evaluation draws; vertical bars mark medians. The plan-pullback bound of Proposition 4 is valid in every cell; the Lipschitz certificate uses the product of spectral norms; the middle row uses a sampled lower bound on the Lipschitz constant and is not a valid certificate.

8 Empirical Study on Tabular Benchmarks

We apply the procedures of Sections 3 to 6 to public tabular benchmarks, using the real data alone.

8.1 Data and Protocol

We analysed nine public data sets: `adult`, `census-income` and `shuttle` from the UCI Machine Learning Repository (Kelly et al., 2026), `bank-marketing` (Moro et al., 2014), `covertypes` (Blackard and Dean, 1999), `higgs` (Baldi et al., 2014), `letter` (Frey and Slate, 1991), `MagicTelescope` (Bock et al., 2004) and `pendigits` (Alimoglu and Alpaydin, 1997). All except `covertypes` were obtained from OpenML (Vanschoren et al., 2014), and `covertypes` through scikit-learn (Pedregosa et al., 2011); the identifiers, versions and checksums of the downloaded tables, and a digest of every encoded matrix, are stored with the results. Each data set is represented by up to two encodings. Encoding E1 keeps the numeric columns; encoding E2 adds the categorical columns as integer codes. Where a data set has no categorical columns, or where the two encodings coincide, as for `covertypes`, one cell is reported, which gives eleven table–encoding cells. For each cell we standardised the columns and drew 8 000 rows at random (seed 0). Floors were computed by exact assignment on the six-point grid $\{125, 250, \dots, 4\,000\}$ with $n_{\text{rep}} = 12$.

Three further candidate data sets were excluded by a screening step that precedes the analysis, because the premise of the floor law fails on them. On `electricity` (Gama et al., 2004), a non-stationary series, the floor does not decrease with n : the median distances are 0.986/1.170/0.967 at $n = 1\,000/2\,000/4\,000$. On `eeg-eye-state` (Kelly et al., 2026), recording artefacts reach 89.4 standard deviations and dominate the tails, and the floor medians are not monotone in n : the median distance jumps from 0.975 at $n = 1\,000$ to 5.32 at $n = 2\,000$. On `MiniBooNE` (Roe et al., 2005), the floor medians are not monotone in n , and a single power law has $R^2 = 0.59$.

For each cell we also report three dimension estimates: the slope estimate of Definition 3, averaged over eight re-splits; the TwoNN estimator of Facco et al. (2017); and the maximum-likelihood estimator of Levina and Bickel (2004) with 10 neighbours, in the form that Pope et al. (2021) applied to image

data, which averages the inverses of the local estimates. Their *spread*, the ratio of the largest to the smallest estimate, indicates disagreement between the geometric and the transport-based view of the data.

8.2 Floor Laws

Table 4 and Figure 5 report the fitted floor laws. On eight of the eleven cells a single power law describes the floor: the coefficient of determination is at least 0.984, the curvature test does not reject, with p -values of 0.112–0.995, and the exponents range from 4.07 to 13.24, all in the dimension-governed regime. The curvature test rejects a single exponent on `adult:E2` and `covertime` ($p = 0.032$ and 0.042), and the floor of `shuttle:E1` admits no power law.

cell	d_{eff}	R^2	curvature p	admissible	$\hat{d} \pm \text{sd}$	spread
<code>adult:E1</code>	4.07	0.984	0.486	yes	4.48 ± 0.96	1.94×
<code>adult:E2</code>	5.91	0.998	0.032	no	5.79 ± 0.31	1.43×
<code>bank-marketing:E1</code>	4.85	0.993	0.955	yes	5.05 ± 1.15	1.20×
<code>bank-marketing:E2</code>	6.55	0.998	0.995	yes	6.20 ± 0.71	1.23×
<code>census-income:E1</code>	5.51	0.992	0.678	yes	5.18 ± 1.44	2.14×
<code>covertime:E1=E2</code>	5.42	0.975	0.042	no	4.64 ± 0.48	1.32×
<code>higgs:E1</code>	13.24	0.999	0.935	yes	13.31 ± 0.51	1.08×
<code>letter:E1</code>	5.34	0.999	0.596	yes	5.25 ± 0.15	1.10×
<code>MagicTelescope:E1</code>	5.59	0.997	0.172	yes	5.48 ± 0.32	1.26×
<code>pendigits:E1</code>	4.57	0.998	0.112	yes	4.83 ± 0.33	1.29×
<code>shuttle:E1</code>	9.45	0.591	—	no	11.62 ± 19.00	7.25×

Table 4: Floor laws of the eleven benchmark cells (8 000 rows, six-point grid over $[125, 4\,000]$, $n_{\text{rep}} = 12$). d_{eff} and R^2 refer to the fit of Definition 2; admissibility is decided by the curvature test of Definition 4; $\hat{d}$ is the slope estimate with its standard deviation over re-splits; spread is the ratio of the largest to the smallest of the three dimension estimates.

The floor of `shuttle:E1` admits no power law at all. Its coefficient of determination is 0.591, the coefficient of variation of its replicates reaches 1.60 at $n = 125$, and its floor medians over the grid,

$$1.38/2.19/2.00/2.27/1.86/1.71,$$

are not monotone: over the lower half of the grid the floor increases with n , so neither the curvature test nor the half-grid prescriptions are defined. No prescription is returned for this cell. Computed nevertheless, the 90% interval of the full-grid prescription spans factors of 53.4×

8.3 Scope of the Sizing Rule

Table 5 reports, for every cell, the agreement between the prescriptions of the two half-grid fits, the curvature p -value and the severity threshold of Definition 5 with its combined interval.

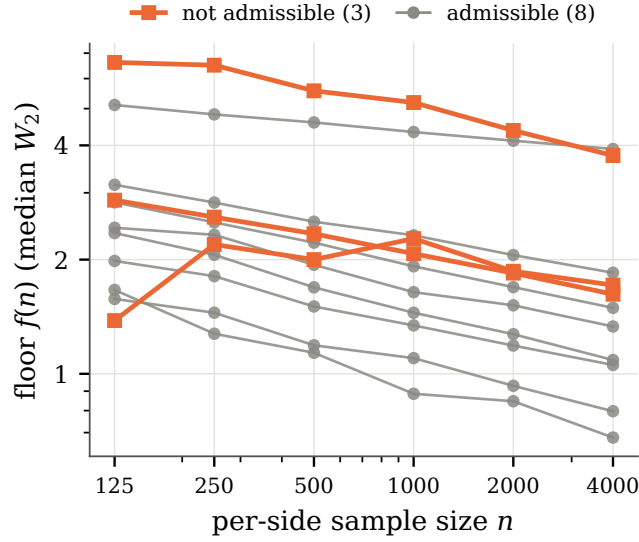


Figure 5: Measured floors $f(n)$ of the eleven benchmark cells on logarithmic axes. Eight cells follow a single power law; the three cells that are not admissible are drawn with squares.

cell	half-fit ratio	curvature p	δ_{50} [90% interval]	observed drift
adult:E1	1.40×	0.486	1.88 [1.72, 2.05]	below
adult:E2	1.51×	0.032	1.19 [1.16, 1.21]	above
bank-marketing:E1	1.36×	0.955	1.48 [1.41, 1.56]	below
bank-marketing:E2	1.18×	0.995	1.20 [1.17, 1.23]	below
census-income:E1	1.35×	0.678	1.86 [1.76, 1.97]	below
covertypes:E1=E2	2.38×	0.042	1.61 [1.54, 1.67]	above
higgs:E1	1.10×	0.935	1.18 [1.17, 1.20]	below
letter:E1	1.18×	0.596	1.13 [1.11, 1.14]	below
MagicTelescope:E1	1.37×	0.172	1.28 [1.24, 1.33]	below
pendigits:E1	1.29×	0.112	1.19 [1.16, 1.21]	below
shuttle:E1	—	—	—	—

Table 5: Scope of the sizing rule. Half-fit ratio: largest ratio between the prescriptions of the low-half and high-half fits. δ_{50} : severity threshold (Definition 5) with its combined binomial and between-seed 90% interval over eight seeds. Last column: position of the observed drift of the cell relative to that interval.

Four results follow. First, the non-rejections bound the undetectable drift at $\delta_{50} \leq 1.88\times$ on every admissible cell and at 1.13 on the most precisely measured one, `letter:E1`. Second, the observed drift of every admissible cell lies below its 50%-detection interval, whereas on the two cells the test rejects, `adult:E2` and `covertime`, it lies above: their curvature is of a size the test detects more than half of the time. Third, on the admissible cells the prescriptions of the two halves of one grid agree within a factor of 1.10–1.40, with a median of 1.32. Fourth, the 90% interval of the prescription is 1.07–2.04 $\times$ wide at the target $n = 10\,000$ and 1.15–2.88 $\times$ at $n = 40\,000$, targets that lie 2.5 $\times$ and 10 $\times$ beyond the largest grid size. Within these limits the sizing rule is identified to within a factor of two to three on these tables. The intervals widen with the distance of the target from the grid, as Proposition 2 predicts.

8.4 The Resolution Atlas

The resolution atlas answers, for each benchmark and sample size and without training any model, the question posed in the introduction: how large must a difference in skill be before an evaluation can detect it?

Definition. For a cell (table, n), let $f(n)$ be the floor and D_{T1} the distance of the independent-columns reference T1 to a fresh real sample of the same size. The difference $S = D_{T1} - f$ is the *skill range*: by Definition 6 a difference ΔD in distance corresponds to a difference $\Delta D/S$ in skill relative to T1. The *resolvable skill difference* is

$$\rho(n) = \frac{(z_{0.975} + z_{0.80}) \sigma_{\text{pair}}}{D_{T1} - f}, \quad (10)$$

the minimum detectable difference of Definition 7 at $\alpha = 0.05$ and power 0.80, expressed as a fraction of the skill range. We measured σ_{pair} directly at the level of T1: in each of 30 replicates we drew a fresh real reference of n rows and two independent samples of T1, and took the standard deviation of the difference of their distances to the reference. A value $\rho = 1$ means that one minimum detectable difference spans the whole range between a generator that discards every dependence between columns and a generator as close to the data as a second real sample. Protocol details are given in Appendix B.7.

Results. Table 6 and Figure 6 report the atlas for eight admissible table–encoding pairs and five sample sizes, $n \in \{250, 500, 1\,000, 2\,000, 4\,000\}$. Over these 40 cells, ρ ranges from 0.014 to 42.124, with a median of 0.236, and 40% of the cells have $\rho \geq 1$.

Resolution varies strongly between benchmarks. At $n = 4\,000$ the resolvable difference spans more than two orders of magnitude. On `letter`, `pendigits` and `MagicTelescope`, a difference of 1.4–2.6% of the skill range is detectable. On `adult:E1` and `bank-marketing:E1`, one minimum detectable difference is at least 1.6 times the whole skill range at every sample size measured, so on these encodings no difference between generators can be resolved with W_2 at these sizes. The ordering follows the amount of dependence that T1 removes: across the 40 cells the Spearman rank correlation between the relative skill range $(D_{T1} - f)/f$ and ρ is -0.85 . Cells of one table are not independent, so the coefficient is descriptive.

In six admissible cells the skill range is smaller than the seed-to-seed standard deviation of D_{T1} : `adult:E1` at $n = 250, 1\,000$ and $2\,000$, and `bank-marketing:E1` at $n = 250, 500$ and $4\,000$. The same holds for two cells that are not admissible, `covertime:E1` at $n = 500$ and `shuttle:E1` at 500

table	$n = 4\,000$			median of ρ over n	proxy / measured SD	cells with span inside noise
	proxy	measured ρ	90% interval			
bank-marketing:E1	8.543	4.488	[3.705, 5.743]	14.760	1.39	3
adult:E1	5.041	3.526	[2.911, 4.512]	3.526	1.43	3
bank-marketing:E2	1.799	0.764	[0.630, 0.977]	1.452	2.36	0
census-income:E1	0.820	0.221	[0.182, 0.283]	0.596	2.58	0
higgs:E1	0.068	0.065	[0.053, 0.083]	0.150	1.75	0
MagicTelescope:E1	0.079	0.026	[0.022, 0.034]	0.056	3.70	0
letter:E1	0.071	0.015	[0.012, 0.019]	0.044	3.23	0
pendigits:E1	0.138	0.014	[0.011, 0.018]	0.037	7.72	0
<i>shuttle:E1</i>	3.419	1.806	[1.491, 2.311]	1.806	2.75	2
<i>covertype:E1</i>	0.955	0.288	[0.238, 0.369]	1.271	1.89	1
<i>adult:E2</i>	0.349	0.150	[0.124, 0.192]	0.237	2.30	0

Table 6: The resolution atlas: the smallest resolvable difference in skill relative to T1, in units of the skill range (Equation (10); $\alpha = 0.05$, 80% power), for $n \in \{250, 500, 1\,000, 2\,000, 4\,000\}$ per side. “Measured” uses the paired standard deviation measured with 30 replicates, with a 90% interval from the χ^2 law of the sample standard deviation; “proxy” uses the floor proxy of Section 5.4. The last column counts the sample sizes at which the skill range is smaller than the seed-to-seed standard deviation of D_{T1} ; the column before it gives the median ratio of the proxy to the measured paired standard deviation. Rows are ordered by the measured value at $n = 4\,000$; the three cells that are not admissible (italic) are shown for completeness and excluded from all summaries.

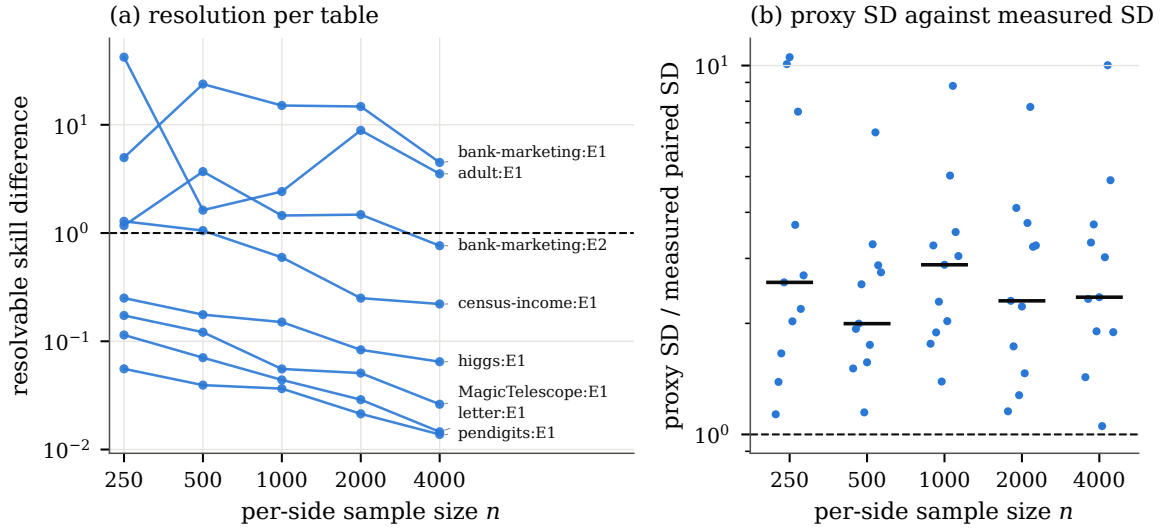


Figure 6: The resolution atlas. (a) Smallest resolvable skill difference ρ against the per-side sample size, one line per admissible table–encoding pair; the dashed line marks $\rho = 1$, where one minimum detectable difference equals the whole skill range. (b) Ratio of the floor proxy to the measured paired standard deviation for all 55 cells; bars mark medians.

and 1 000. At these sizes an evaluation cannot detect that a generator has removed every dependence in the table. For `adult:E1` the reason is visible in the data: its six numeric columns have a largest pairwise absolute correlation of 0.155 and a median of 0.057, so there is little dependence to remove, and removing it changes W_2 very little.

The floor proxy of Section 5.4 exceeds the measured paired standard deviation in all 55 cells, by factors between 1.05 and 10.53 with a median of 2.36 (Figure 6b). It is therefore a conservative substitute when the direct measurement is too expensive, and it can overstate the required difference several-fold.

Scope. The atlas concerns W_2 at $n \leq 4\,000$ per side. Marginal statistics, downstream-task scores and kernel tests can have higher power against specific departures; W_2 is an omnibus joint distance, and omnibus tests lose power as the dimension grows. Encoding E1 is numeric only and E2 embeds categorical columns as integers, so neither is a canonical representation of a mixed-type table (Section 10). The paired standard deviation of a single cell rests on 30 replicates and is determined to within roughly -17% and $+28\%$; conclusions are drawn from the pattern across cells. Finally, every quantity here holds the generator fixed and varies only the samples. Retraining a generator adds variability, which Dufour et al. (2026) found to be $3.2\times$ larger than the resampling variability for image models, so the atlas gives lower bounds on the differences required when trained models are compared.

8.5 Certification Margins

Every admissible cell has a measured dimension of at most 13.31, below the certification limits of 26.0 at $N = 2\,048$ and 29.6 at $N = 4\,096$ that bracket the 4 000 observations per side of these tables. To quantify the margin, we propagated three sources of uncertainty jointly by Monte Carlo with 8 000 draws per table: the sampling uncertainty of $\hat{d}$ estimated by subsampling, a law-to-law residual in r resampled from the transfer study of Section 6.5, and the sampling error of the calibration map of Section 6.4 (Appendix B.6). The margin in dimension units is the ratio of the crossing to the drawn dimension, and the margin in r units is $1/r$ at the drawn dimension, the quantity that decides vacuity. Table 7 reports the fifth percentiles for the eight admissible cells.

The fifth-percentile margins in r units range from $1.30\times$ to $3.23\times$, and no draw was vacuous. Margins in dimension units are larger because the crossing responds weakly to r (Section 6.3); the r units are the relevant ones. Eight of the eleven cells are flagged as heavy-tailed, and the last column therefore charges the $+44.1\%$ residual of the Student- t law in full. Seven cells keep a margin above one, between $1.52\times$ and $2.24\times$. The exception is `higgs:E1`, whose measured dimension of 13.31 is the largest of the study and whose tail index of 3.65 is below the threshold: if its floor ratio carried the full heavy-tail error, its fifth-percentile margin would be $0.90\times$ and a certificate could not be ruled informative. With this exception, the statistical term does not rule out certificates of the form Equation (6) on these tables, whose size of 4 000 observations per side lies inside the range of the reference tables.

9 Implementation and Verification

All procedures of this paper are implemented in the Python package `certgen` (version 3.3.1), which depends only on NumPy (Harris et al., 2020) and SciPy (Virtanen et al., 2020) and computes every distance by exact assignment. The package returns no sample size when a floor profile is not

cell	$\hat{d}$	$q_{0.05}$, dimension units	$q_{0.05}$, r units	$q_{0.05}$, r units, heavy tails
pendigits:E1	4.83	6.92×	3.23×	2.24×
letter:E1	5.25	6.40×	3.08×	2.13×
MagicTelescope:E1	5.48	5.92×	2.70×	1.87×
bank-marketing:E1	5.05	5.91×	2.54×	1.77×
adult:E1	4.48	5.84×	2.45×	1.70×
census-income:E1	5.18	5.47×	2.32×	1.61×
bank-marketing:E2	6.20	4.98×	2.18×	1.52×
higgs:E1	13.31	2.13×	1.30×	0.90×

Table 7: Certification margins of the eight admissible cells with three sources of uncertainty propagated jointly (8 000 Monte-Carlo draws per cell, reference tables evaluated at the cells’ own sample size). No draw was vacuous in any cell, which bounds the probability of vacuity by about 3/8 000 at 95% confidence (Hanley and Lippman-Hand, 1983). The last column divides the r -unit margin by 1.441, which charges the largest heavy-tail residual of Section 6.5 in full.

admissible or when $d_{\text{eff}} < 4$, reporting the range of half-fit prescriptions instead; it applies the interval calibration of Section 4.6 only near the regime in which the calibration was measured; and it reports every detectability statement together with its significance level and power.

Dimension estimators. The TwoNN and maximum-likelihood estimators use nearest-neighbour distances and are undefined when observations coincide. We exclude zero distances instead of replacing them by a small constant: on a quantised Gaussian copula with a single true exponent, replacing zero distances by 10^{-12} produces a spurious 4.80× spread between the estimators. The fraction of tied observations is reported with every estimate. The confidence interval for the slope estimate, obtained from the mean of eight re-split replicates of the exponent, uses the Student- t quantile with seven degrees of freedom; the normal quantile delivers a coverage of 0.862 at nominal 0.90 in simulation, the t quantile 0.902. We implement the three estimators directly because the slope estimate reuses the replicate structure of the floor computation; the package `skdim` (Bac et al., 2021) provides independent implementations of the geometric estimators.

Computational cost. On one processor core, the floor study of Section 8.2 takes about 26 minutes for the eleven cells, without the download of the data; the resolution atlas of Section 8.4 takes about 1.8 CPU hours, most of it in the paired replicates at $n = 4\,000$; and the reference tables of Section 6.3 take about 1.4 CPU hours up to $N = 8\,192$ and 46 CPU hours at $N = 16\,384$ and $32\,768$. Table 8 in Appendix B.2 gives the cost of a single exact solve.

Reproducibility. Every experiment writes its results to a file in the software repository, every number and figure in this paper is computed from those files, and a script re-executes every experiment and compares the regenerated results with the stored ones. The experiments that download public data sets record the preprocessing described in Section 8.1.

Machine-checked statements. The formal statements of this paper are proved in the Lean 4 proof assistant (de Moura and Ullrich, 2021). A development without external dependencies proves, over an abstract ordered field whose axioms are listed explicitly, Propositions 2 to 4 and Lemma 1; the

deterministic core of Proposition 1, namely that a metric discrepancy between distinct empirical measures is positive and that replicate values with mean zero, not all zero, include a negative value; the validity of both Monte-Carlo p -values of Section 3.2 and the resulting requirement $n_{\text{rep}} \geq 19$; the relation between the multiplier of Equation (5) and power; the bounds used for the severity threshold; and the invariance and limits of the skill score of Definition 6. The axioms used by 51 exported results are checked automatically, so an incomplete proof fails the build. A second development instantiates the ordered field by the real numbers of the Lean mathematical library (The mathlib Community, 2020) and proves there the reduction of Section 3.5, that the transport problem between two uniform empirical measures of equal size has an optimal solution given by an assignment, using the Birkhoff–von Neumann theorem (Birkhoff, 1946). The measurements of this paper, such as fitted exponents, coverages and tightness ratios, are empirical and are not formalised.

Availability. The software, the stored results, the figure scripts and the Lean developments are available at <https://github.com/kahyaoglu20/certgen>.

10 Discussion

10.1 Mixed-Type Data

Tabular data usually contain categorical and mixed-type columns, and a Euclidean distance such as W_2 is defined only after the rows have been embedded in $\mathbb{R}^P$. No embedding is canonical, and introducing a new type-aware distance would add an unvalidated component to a framework whose purpose is to measure the resolution of distances already in use. We therefore recommend a sensitivity analysis: fix at least two embeddings before any generator is evaluated, apply the full procedure under each, and report where the conclusions differ. A conclusion that changes with the embedding is a statement about the embedding. Integer codes of categorical columns produce lattice structure, whose Wasserstein rate is that of a discrete measure, and ties affect nearest-neighbour estimators. The effect is visible in Table 4: on `adult`, the spread of the three dimension estimators is $2.13\times$ under E1 and $1.52\times$ under E2.

10.2 Limitations

1. *Range of the certification tables.* The reference tables cover $N \leq 32\,768$, whereas many tabular evaluations use 10^4 to 10^5 observations. At $N = 65\,536$ the dense cost matrix of one exact solve alone occupies about 34 GB of memory. Beyond the size that exact solves reach on the available memory, extending the tables requires an approximate transport estimator whose bias has been measured against exact transport on the range where both are available.
2. *Heavy tails.* The reference tables do not transfer to laws with polynomial tails, and eight of the eleven benchmark cells are flagged by the tail diagnostic. A reference table for heavy-tailed laws would remove the need to charge the largest observed residual in full.
3. *Power of the curvature test.* At the floor precision of the benchmark tables the test has power between 0.08 and 0.885 against the two-scale alternative, so eight non-rejections are compatible with departures from a single exponent that the test cannot see. Measures whose effective dimension changes with scale (Weed and Bach, 2019) are a principled source of such departures. The severity thresholds of Table 5 state which departures are excluded.

4. *Sampling variability only.* The floor and the atlas vary the samples and hold the generator fixed, so they give lower bounds on the resolution required to compare trained models (Bouthillier et al., 2021).
5. *Screening quantities.* The minimum detectable difference is a design and screening quantity. Formal comparisons of two models against a shared reference should use the relative-similarity test of Bounliphone et al. (2016) or its relative Kullback–Leibler analogue (Gao et al., 2025).
6. *Leakage detection.* A memoriser that perturbs records by more than about $0.44\times$ the median distance between records escapes the screen, and on densely sampled finite supports memorisation cannot be identified by any detector.
7. *Size of the data.* The fitting grid stops at $n = N/2$, which limits both the power of the curvature test and the distance over which a prescription extrapolates; every extrapolation in this paper states its distance beyond the grid.
8. *Calibration of the interval.* The calibrated level 0.9125 delivers 0.90 as a median over its regime; the coverage of individual cells varies between 0.860 and 0.920 at nominal 0.90.

10.3 Scope

The framework does not introduce a new discrepancy: it uses W_2 for the reason given by Proposition 1. It ranks no generators; the resolution atlas establishes its conclusions without training one. Leakage screening serves the validity of fidelity statements and provides no privacy guarantee. Finally, the certification limit concerns what can be proved about a generator at a given sample size. A generator may be much better than any certificate available at that size can show.

11 Conclusion

The distance between a real and a synthetic sample is a measurement, and its resolution can be determined before the measurement is made. From the real data alone, the resolution floor gives the distance that an ideal generator would still show, the smallest difference that two models must exhibit to be distinguished, and the dimension above which a latent-generative certificate cannot be informative. The sample-size rule derived from the floor rests on a fitted exponent, and we accompany it with a test of the single-exponent assumption whose size, power and severity are measured. On standard tabular benchmarks, the resolvable difference varies by more than two orders of magnitude between data sets at the same sample size. It therefore has to be computed for each evaluation, which requires minutes to hours of computation on the real data and no trained model.

Acknowledgments and Disclosure of Funding

This work received no specific funding. The author declares no competing interests.

Appendix A. Proofs

Proof of Proposition 1

Let μ_n and μ'_n be the empirical measures of X_n and X'_n . Since W_2 is a metric on the probability measures with finite second moment, $W_2(\mu_n, \mu'_n) \geq 0$, with equality if and only if $\mu_n = \mu'_n$, that is, if and only if the two samples coincide as multisets. Under Assumption A2 the probability that any two of the $2n$ draws coincide is zero, so $\mathbb{P}(\mu_n = \mu'_n) = 0$ and $W_2(\mu_n, \mu'_n) > 0$ almost surely. A non-negative random variable that is almost surely positive has positive expectation, so $\mathbb{E}[D(X_n, X'_n)] > 0$. The equality $W_2(P, P) = 0$ follows from the diagonal coupling.

For the second statement, suppose $\mathbb{E}[\hat{D}] = 0$ and $\hat{D}$ is not almost surely constant. If $\hat{D} \geq 0$ held almost surely, then $\mathbb{E}[\hat{D}] = 0$ would force $\hat{D} = 0$ almost surely, a contradiction. Hence $\hat{D} < 0$ with positive probability, and on that event $\max(\hat{D}, 0) = 0$ and $\log \max(\hat{D}, 0)$ is undefined. $\square$

Remark 1. The conclusion $\mathbb{E}[D] > 0$ requires only $\mathbb{P}(\mu_n = \mu'_n) < 1$, which is weaker than Assumption A2. For laws with atoms the probability of coincident observations is positive and the floor is correspondingly smaller. This is why coarse lattices, such as integer-coded categorical columns, need separate treatment (Sections 6.5 and 10.1).

Proof of Proposition 2

Taking logarithms in Equation (2) gives $\log n^* = d_{\text{eff}}(\log C - \log \phi)$, which is differentiable on $\{\phi < C\}$. Differentiation with respect to $\log C$ gives d_{eff} , and differentiation with respect to d_{eff} gives $\log(C/\phi)$, which equals $\log n^*/d_{\text{eff}}$ by the same relation. Replacing C by $C(1 + \varepsilon)$ multiplies n^* by $(1 + \varepsilon)^{d_{\text{eff}}}$, which equals $e^{d_{\text{eff}}\varepsilon}$ to first order. Replacing d_{eff} by $d_{\text{eff}}(1 + \delta)$ multiplies $\log n^*$ by $1 + \delta$ and hence n^* by $(n^*)^\delta$. $\square$

Proof of Proposition 3

Write the right-hand side of Equation (6) as $S + E + T$ with $E = W_2(Y_N, \tilde{Y}_M) \geq 0$ and $T = W_2(\tilde{Y}_M, \tilde{Y}) \geq 0$. If $S > \tau$, then $S + E + T \geq S > \tau$, so the bound exceeds the trivial bound τ . $\square$

Remark 2. The antecedent of Proposition 3 is the statistical term S and not the surrogate r of Equation (7). The triangle inequality bounds the two-sample distance by the sum of two one-sample distances and therefore yields only $\mathbb{E} S \geq r\tau/2$. The ratio between the two quantities is measured in Appendix B.3.

Proof of Lemma 1 and Proposition 4

Let $\pi \in \Pi(\Sigma, \Sigma')$ and let $\rho = (g \otimes g)_\# \pi$ be the pushforward of π under $(z, z') \mapsto (g(z), g(z'))$. For a Borel set $A \subseteq \mathbb{R}^m$,

$$\rho(A \times \mathbb{R}^m) = \pi(g^{-1}(A) \times \mathcal{Z}) = \Sigma(g^{-1}(A)) = (g_\# \Sigma)(A),$$

where $\mathcal{Z}$ is the latent space and the second equality uses that the first marginal of π is Σ ; the second marginal is treated in the same way. Hence $\rho \in \Pi(g_\# \Sigma, g_\# \Sigma')$, and since W_2^2 is the infimum of the transport cost over all couplings,

$$W_2(g_\# \Sigma, g_\# \Sigma')^2 \leq \int \|u - v\|^2 d\rho(u, v) = \int \|g(z) - g(z')\|^2 d\pi(z, z'),$$

by the change-of-variables formula for pushforward measures. The integrability assumptions make the right-hand side finite.

For the proposition, take $\Sigma = \Sigma_N$ and $\Sigma' = \Sigma'_N$. The assignment σ^* induces the coupling $\pi^* = N^{-1} \sum_i \delta_{(z_i, z'_{\sigma^*(i)})}$, which lies in $\Pi(\Sigma_N, \Sigma'_N)$ because σ^* is a bijection. Substituting π^* into Equation (8) gives Equation (9). Its right-hand side requires $g(z_i)$ and $g(z'_j)$ for all i and j , that is $2N$ evaluations of g , and no derivative of g . $\square$

Remark 3. Any bijection yields a valid coupling and hence a valid bound; optimality of σ^* in the latent space serves tightness. The Lipschitz certificate is one way of bounding the integrand of Equation (8); evaluating the integrand directly is another.

Appendix B. Experimental Details

B.1 Synthetic Laws

Gaussian copula. A d -dimensional Gaussian copula with correlation $\rho^{|i-j|}$ between coordinates i and j , $\rho = 0.8$, and bounded, non-Gaussian marginals: with $U_j = \Phi(Z_j)$ for the correlated Gaussian vector Z , even coordinates are $\exp(2U_j)$ and odd coordinates are $\tan(0.6\pi(U_j - 1/2))$. The six- and eight-dimensional versions are used in Sections 3 to 5; where stated, columns are standardised.

Two-scale floor. The alternative of Sections 4.2 and 4.4 is the floor profile $f(n) = n^{-1/2} + 0.35 n^{-1/8}$, whose local exponent moves from about 2 at small n towards 8 at large n ; replicate noise is added at the stated precision. The null profiles of the power study are single power laws $n^{-0.2}$, that is, $d_{\text{eff}} = 5$.

Isotropic Gaussians. Standard normal laws in $\mathbb{R}^d$, for which intrinsic and ambient dimension coincide and $\tau = \sqrt{d}$; they are used for the size study of Table 1, the reference tables and the dimension calibration.

B.2 Cost of Exact Transport

Table 8 reports the median wall time of one exact two-sample distance on standard normal samples: a dense squared-Euclidean cost matrix followed by an exact linear assignment, on one core of an Intel Core i7-14700HX processor. The fitted growth exponents in n for $n \geq 1\,024$ are 2.56 for $d = 8$ and 2.45 for $d = 32$. The cost matrix occupies $8n^2$ bytes.

n per side	512	1 024	2 048	4 096	8 192
$d = 8$, seconds	0.012	0.060	0.36	2.1	12
$d = 32$, seconds	0.013	0.066	0.37	2.1	11

Table 8: Wall time of one exact two-sample W_2 (median of repeated solves).

B.3 Two-Sample and One-Sample Distances

To relate the surrogate of Equation (7) to the statistical term, we compared, on isotropic Gaussians with $d \in \{2, 4, 8, 16, 32\}$ and $N \in \{256, 1\,024\}$, the two-sample distance $D(X_N, X'_N)$ with the distance between X_N and a reference sample of 2 000 observations, which stands in for the population. The ratio of their means lies between 1.007 and 1.16; it decreases with d and with N and is below the heuristic value $2^{1/d}$ in every cell.

B.4 Reference Tables and the Calibration Map

Each cell of Table 3 is the mean over 20 replicates of $D(X_N, X'_N)/\sqrt{d}$ for two independent standard normal samples of N observations in $\mathbb{R}^d$. Each cell of Table 9 is the mean over 20 replicates of the slope estimate of Definition 3 on a standard normal sample of N observations. The crossing in nominal units is the largest d with $r(d, N) < 1$, obtained by interpolation of $\log r$ in $\log d$; it is mapped to measured units by interpolation of Table 9. The 90% intervals of Section 6.4 are quantiles over 2 000 recomputations of the crossing in which every cell of both tables is perturbed by its standard error and the perturbed tables are made monotone in d .

At $N = 16\,384$ and $32\,768$ one solve holds a dense cost matrix of $8N^2$ bytes, so the replicates ran as separate processes admitted under a memory budget, with the same estimator, sample construction and random seeds as at smaller N . One transport solve took a median of 123 seconds and at most 2.3 GB of memory at $N = 16\,384$, and 492 seconds and 8.8 GB at $N = 32\,768$; one slope estimate at $N = 32\,768$ took a median of 274 seconds, on the processor of Appendix B.2. Before these solves, the same procedure recomputed six cells of the $N = 8\,192$ tables, and all six values agreed with the tables bit for bit.

nominal d	2	4	8	16	32	64	128
$N = 256$	2.8	4.3	7.6	12.5	19.9	31.4	48.7
$N = 1\,024$	2.7	4.4	7.5	13.1	21.6	35.2	53.8
$N = 2\,048$	2.5	4.2	7.7	13.5	22.2	36.4	56.4
$N = 4\,096$	2.6	4.2	7.8	13.7	23.2	38.3	59.3
$N = 8\,192$	2.4	4.2	7.7	13.9	23.9	38.5	60.9
$N = 16\,384$	2.5	4.1	7.9	14.3	24.4	39.7	62.9
$N = 32\,768$	2.4	4.2	7.9	14.4	25.0	41.0	65.5

Table 9: Calibration map: mean slope dimension estimate $\hat{d}$ of isotropic Gaussian samples of nominal dimension d and size N (20 replicates per cell).

B.5 Transfer Study

All laws live in $\mathbb{R}^{64}$. The anisotropic Gaussians have covariance eigenvalues decaying geometrically with ratio 0.95, 0.90 or 0.85 in a randomly rotated basis. The factor models have $k \in \{20, 28\}$ Gaussian factors with a random loading matrix and isotropic noise of standard deviation 0.25. The curved manifolds are images of $m \in \{16, 24, 32\}$ dimensional Gaussian latent variables under a fixed smooth random embedding. The geometry of each law is drawn once and held fixed, so that the two samples of every comparison come from the same law; a comparison of the two sample covariances guards against violations. Each law is sampled at $N \in \{256, 1\,024\}$, and residuals are reported for the band of measured dimensions in which the crossing lies. The probe laws of the second experiment are a multivariate Student- t law with $\nu \in \{3, 5\}$ and the anisotropic scale above, a rotated Laplace law, the factor models rounded to a grid of step 0.5 or 1 in every coordinate, and the factor models with half of the coordinates rounded.

B.6 Margin Analysis

For each cell and each of 8 000 Monte-Carlo draws we sample the dimension from a normal law with the measured mean of $\hat{d}$ and its subsampling standard deviation; convert it to nominal units through the calibration map, whose cells are perturbed by their standard errors; draw a law-to-law residual in r from the in-band cells of the transfer study; and evaluate the reference tables at the cell’s own sample size of 4 000 observations per side. The margin in dimension units is the resulting crossing divided by the drawn nominal dimension, and the margin in r units is $1/r$ at the drawn nominal dimension, adjusted by the residual.

B.7 Resolution Atlas

For each table and encoding we used up to 20 000 rows. At each sample size, the floor is the median of $n_{\text{rep}} = 20$ replicates, and D_{T1} is the mean over five independent samples of T1, each compared with a fresh real sample. For the paired standard deviation, each of 30 replicates draws a fresh real reference of n rows and two independent samples of T1 and records the difference of their distances to the reference; the 90% interval of the standard deviation follows from the χ^2 law with 29 degrees of freedom. The skill range lies inside the baseline noise when $D_{\text{T1}} - f$ is smaller than the standard deviation of D_{T1} across the five samples. The floor proxy of Table 6 uses the same cells with σ_{pair} replaced by $\sqrt{2} \text{CV}(f) D_{\text{T1}}$.

B.8 Variational Autoencoders

The `digits` data set is standardised on the training split of 1 300 rows. The two architectures are multilayer perceptrons with rectified linear activations: latent dimension 8 with two hidden layers of width 128, and latent dimension 12 with three hidden layers of width 256, mirrored in encoder and decoder. Each model is trained in PyTorch (Paszke et al., 2019) with Adam (Kingma and Ba, 2015) at learning rate 10^{-3} for 150 epochs with three seeds. The Lipschitz certificate multiplies the spectral norms of the decoder’s weight matrices, the activations being 1-Lipschitz, and multiplies the result by the latent W_2 distance. The sampled lower bound is the largest ratio $\|g(z) - g(z')\|/\|z - z'\|$ over sampled latent pairs.

References

- M. Ajtai, J. Komlós, and G. Tusnády. On optimal matchings. *Combinatorica*, 4(4):259–264, 1984. doi: 10.1007/BF02579135.
- F. Alimoglu and E. Alpaydin. Combining multiple representations and classifiers for pen-based handwritten digit recognition. In *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, volume 2, pages 637–640, 1997. doi: 10.1109/ICDAR.1997.620583.
- J. Bac, E. M. Mirkes, A. N. Gorban, I. Tyukin, and A. Zinovyev. Scikit-dimension: a Python package for intrinsic dimension estimation. *Entropy*, 23(10):1368, 2021. doi: 10.3390/e23101368.
- P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5:4308, 2014. doi: 10.1038/ncomms5308.

- L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004. doi: 10.1016/S0047-259X(03)00079-4.
- J. Bigot, E. Cazelles, and N. Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2): 5120–5150, 2019. doi: 10.1214/19-EJS1637.
- G. Birkhoff. Tres observaciones sobre el algebra lineal. *Universidad Nacional de Tucumán, Revista Serie A*, 5:147–151, 1946.
- J. A. Blackard and D. J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 1999. doi: 10.1016/S0168-1699(99)00046-0.
- A. Block, Z. Jia, Y. Polyanskiy, and A. Rakhlin. Intrinsic dimension estimation using Wasserstein distance. *Journal of Machine Learning Research*, 23(313):1–37, 2022.
- S. Bobkov and M. Ledoux. One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Memoirs of the American Mathematical Society*, 261(1259), 2019. doi: 10.1090/memo/1259.
- R. K. Bock, A. Chilingarian, M. Gaug, F. Hakl, T. Hengstebeck, M. Jiřina, J. Klaschka, E. Kotrč, P. Savický, S. Towers, A. Vaiciulis, and W. Wittek. Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research Section A*, 516(2–3):511–528, 2004. doi: 10.1016/j.nima.2003.08.157.
- N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. doi: 10.1007/s10851-014-0506-3.
- V. Borisov, K. Seßler, T. Leemann, M. Pawelczyk, and G. Kasneci. Language models are realistic tabular data generators. In *International Conference on Learning Representations*, 2023.
- W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton. A test of relative similarity for model selection in generative models. In *International Conference on Learning Representations*, 2016.
- X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Sepah, E. Raff, K. Madan, V. Voleti, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, 2021.
- D. Card, P. Henderson, U. Khandelwal, R. Jia, K. Mahowald, and D. Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9263–9274, 2020. doi: 10.18653/v1/2020.emnlp-main.745.
- G. Carlier, A. Figalli, Q. Mérigot, and Y. Wang. Sharp comparisons between sliced and standard 1-Wasserstein distances. arXiv:2510.16465, 2025.

- H. Chen. What FID hides: detecting, ranking, and diagnosing deviations in generative evaluation. arXiv:2608.24881, 2026.
- M. J. Chong and D. Forsyth. Effectively unbiased FID and Inception Score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6069–6078, 2020. doi: 10.1109/CVPR42600.2020.00611.
- A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. doi: 10.1137/070710111.
- D. F. Crouse. On implementing 2D rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016. doi: 10.1109/TAES.2016.140952.
- M. Cuturi. Sinkhorn distances: lightspeed computation of optimal transportation distances. In *Advances in Neural Information Processing Systems*, volume 26, 2013.
- Q. Dao, H. Phung, B. Nguyen, and A. Tran. Flow matching in latent space. arXiv:2307.08698, 2023.
- A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, 1997. doi: 10.1017/CBO9780511802843.
- L. de Moura and S. Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, volume 12699 of *Lecture Notes in Computer Science*, pages 625–635. Springer, 2021. doi: 10.1007/978-3-030-79876-5_37.
- E. del Barrio and J.-M. Loubes. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951, 2019. doi: 10.1214/18-AOP1275.
- E. del Barrio, E. Giné, and C. Matrán. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *The Annals of Probability*, 27(2), 1999. doi: 10.1214/aop/1022677394.
- R. M. Dudley. The speed of mean Glivenko–Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969. doi: 10.1214/aoms/1177697802.
- N. Dufour, A. A. Efros, and P. Pérez. The FID lottery: quantifying hidden randomness in generative-model evaluation. arXiv:2606.20536, 2026.
- G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017.
- B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.
- C. Esteban, S. L. Hyland, and G. Rätsch. Real-valued (medical) time series generation with recurrent conditional GANs. arXiv:1706.02633, 2017.
- E. Facco, M. d’Errico, A. Rodriguez, and A. Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017. doi: 10.1038/s41598-017-11873-y.

- M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738, 2015. doi: 10.1007/s00440-014-0583-7.
- P. W. Frey and D. J. Slate. Letter recognition using Holland-style adaptive classifiers. *Machine Learning*, 6(2):161–182, 1991. doi: 10.1007/BF00114162.
- J. H. Friedman. On multivariate goodness-of-fit and two-sample testing. In *Proceedings of PHYSTAT 2003, eConf C030908*, pages 311–313, 2003.
- J. Gama, P. Medas, G. Castillo, and P. Rodrigues. Learning with drift detection. In *Advances in Artificial Intelligence – SBIA 2004*, volume 3171 of *Lecture Notes in Computer Science*, pages 286–295. Springer, 2004. doi: 10.1007/978-3-540-28645-5_29.
- Z. Gao, Y. Sun, and H. Su. Statistical inference for generative model comparison. arXiv:2501.18897, 2025.
- A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19, pages 513–520, 2007.
- A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.
- J. A. Hanley and A. Lippman-Hand. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA*, 249(13):1743–1745, 1983. doi: 10.1001/jama.1983.03330370053031.
- C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2.
- M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975. doi: 10.1214/aos/1176343247.
- A. C. A. Hope. A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society, Series B*, 30(3):582–598, 1968. doi: 10.1111/j.2517-6161.1968.tb00759.x.
- S. Hundrieser, T. Staudt, and A. Munk. Empirical optimal transport between different measures adapts to lower complexity. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 60(2):824–846, 2024. doi: 10.1214/23-AIHP1369.

- S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar. Rethinking FID: towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024. doi: 10.1109/CVPR52733.2024.00889.
- R. Jonker and A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987. doi: 10.1007/BF02278710.
- L. V. Kantorovich. On the translocation of masses. *Doklady Akademii Nauk SSSR*, 37:199–201, 1942. English translation: *Journal of Mathematical Sciences* 133(4):1381–1382, 2006.
- M. Kelly, R. Longjohn, and K. Nottingham. The UCI machine learning repository. <https://archive.ics.uci.edu>, 2026. Accessed 13 October 2026.
- D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.
- H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.
- T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- W. Lee, R. C. W. O’Neill, D. Zou, J. Calder, and G. Lerman. Geometry-preserving encoder/decoder in latent generative models. arXiv:2501.09876, 2025.
- E. Levina and P. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, volume 17, 2004.
- D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.
- E. Miller. Adding error bars to evals: a statistical approach to language model evaluations. arXiv:2411.00640, 2024.
- S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014. doi: 10.1016/j.dss.2014.03.001.
- T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019. doi: 10.1002/sim.8086.
- J. Niles-Weed and P. Rigollet. Estimation of Wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022. doi: 10.3150/21-BEJ1433.
- J. Niles-Weed and J. Shkrob. Nearly sharp comparison results for sliced and max-sliced Wasserstein distances. arXiv:2608.13374, 2026.

- N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083, 2018. doi: 10.14778/3231751.3231757.
- A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- N. Patki, R. Wedge, and K. Veeramachaneni. The Synthetic Data Vault. In *IEEE International Conference on Data Science and Advanced Analytics*, pages 399–410, 2016. doi: 10.1109/DSAA.2016.49.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.
- G. Peyré and M. Cuturi. Computational optimal transport: with applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/22000000073.
- P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.
- J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, volume 6667 of *Lecture Notes in Computer Science*, pages 435–446. Springer, 2012. doi: 10.1007/978-3-642-24785-9_37.
- A. Ramdas, N. García Trillos, and M. Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017. doi: 10.3390/e19020047.
- B. P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu, and G. McGregor. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A*, 543(2–3):577–584, 2005. doi: 10.1016/j.nima.2004.12.018.
- M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
- T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, pages 2234–2242, 2016.
- A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton. MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81, 2023.
- D. J. Schuurmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419.
- M. Sommerfeld and A. Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society, Series B*, 80(1):219–238, 2018. doi: 10.1111/rssb.12236.

- C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.
- G. J. Székely and M. L. Rizzo. Energy statistics: a class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.
- C. Taming, M. Sommerfeld, and A. Munk. Empirical optimal transport on countable metric spaces: distributional limits and statistical applications. *The Annals of Applied Probability*, 29(5): 2744–2781, 2019. doi: 10.1214/19-AAP1463.
- The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, pages 367–381, 2020. doi: 10.1145/3372885.3373824.
- B. van Breugel, H. Sun, Z. Qian, and M. van der Schaar. Membership inference attacks against synthetic data through overfitting detection. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 3493–3514, 2023.
- J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014. doi: 10.1145/2641190.2641198.
- C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. doi: 10.1007/978-3-540-71050-9.
- A. Virmaux and K. Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
- P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.
- J. von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. In *Contributions to the Theory of Games II*, number 28 in *Annals of Mathematics Studies*, pages 5–12. Princeton University Press, 1953. doi: 10.1515/9781400881970-002.
- 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. doi: 10.3150/18-BEJ1065.
- W. J. Westlake. Response to T. B. L. Kirkwood: bioequivalence testing—a need to rethink. *Biometrics*, 37(3):589–594, 1981. doi: 10.2307/2530573.
- L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- S. C.-H. Yang, B. Eaves, M. Schmidt, K. Swanson, and P. Shafto. Structured evaluation of synthetic tabular data. arXiv:2403.10424, 2024.