Nearest-Neighbor Radii under Dependent Sampling

Yuanyuan Gao[*]
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720

Yilong Hou[*]
Department of Biostatistics
University of California, Berkeley
Berkeley, CA 94720

Zhexiao Lin[*,†]
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720

Abstract

Nearest-neighbor methods are fundamental to classical and modern machine learning, yet their geometric properties are typically analyzed under independent sampling. In this paper, we study the nearest-neighbor radii under dependent sampling. We consider strong mixing dependent observations and ask whether dependence changes the scale of nearest-neighbor neighborhoods. We establish distribution-free almost sure convergence under polynomial mixing and sharp non-asymptotic moment bounds under geometric mixing. The moment bounds depend on the local intrinsic dimension rather than the ambient dimension, making the results applicable to high-dimensional data supported on lower-dimensional manifolds. Synthetic experiments and real-world time-series benchmarks illustrate the theory, showing that nearest-neighbor geometry remains informative under dependent sampling.

1 Introduction

Nearest-neighbor (NN) methods remain one of the most fundamental ideas in machine learning. They underlie classical procedures such as k-nearest-neighbor classification [13] and regression [56], nonparametric density estimation and local averaging [45], and they also appear implicitly in a much broader range of modern techniques built on local neighborhoods, including graph-based semi-supervised learning [65], manifold methods [6], information retrieval [58], and language modeling [32, 60, 61]. In all of these methods, a basic geometric quantity at the center of the analysis is the distance from a query point x to its kth nearest neighbor in the sample. This random radius determines the scale of the local neighborhood around x, and therefore controls how much of the ambient space the method actually "sees."

From a statistical perspective, nearest-neighbor radii determine the central bias–variance tradeoff of local methods [56, 7]. If the radius is too large, a local estimator averages over points that are geometrically far from the query and thus introduces bias. If the radius is too small, the estimator uses too few effective observations and becomes unstable. In density estimation, the k-NN radius is the random bandwidth that determines the volume of the adaptive neighborhood. In regression and classification, it controls the locality of smoothing. In manifold and graph-based methods, it determines neighborhood connectivity and the scale at which local geometry is resolved.

[*]Authors are listed in alphabetical order.

[†]Corresponding author: zhexiaolin@berkeley.edu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Understanding the scaling of the k-NN radius is therefore a prerequisite for understanding the performance of a wide class of machine learning algorithms.

In the classical i.i.d. setting, the behavior of k-NN radius is well understood: under standard local mass conditions, the canonical scale is $(k/n)^{1/d}$, or more generally, $(k/n)^{1/s}$ when the local geometry is determined by an intrinsic dimension s [45, 7, 35]. However, many learning problems do not arise from independent samples. Data are often collected sequentially, spatially, or through interaction with a dynamical environment [50, 57]. In such settings, observations can be strongly dependent even when their marginal distribution is stable. This raises a basic question:

How does dependence affect the geometry of nearest-neighbor neighborhoods?

This question is both mathematically natural and practically important. A large body of machine learning theory relies on i.i.d. nearest-neighbor geometry either explicitly or implicitly. Yet once dependence is present, standard concentration arguments no longer apply directly, and even basic geometric quantities such as the k-NN radius become more delicate. If dependence were to fundamentally change the scaling of nearest-neighbor radii, then the geometry underlying many local learning procedures would also change. On the other hand, if the canonical radius scale remains unchanged under weak dependence, this would help justify the use of nearest-neighbor and local methods beyond the i.i.d. regime.

In this paper, we study the nearest-neighbor radii under dependent sampling. We consider a triangular array of $\mathbb{R}^d$-valued observations with a common marginal distribution and strong mixing dependence along each row. This framework is general enough to cover the usual stationary-sequence setting as a special case, while also allowing the dependence structure to vary with n. Our object of study is the radius $R_{n,k}(x) := \|\mathcal{X}^n_{(k)}(x) - x\|$, that is, the distance from a fixed query point x to its kth nearest neighbor among the n observations in the row.

Our first result is almost sure convergence: under polynomial α-mixing, we show that the nearest-neighbor radius converges almost surely to the distance from x to the support of the marginal law. This gives a dependent-sampling analogue of the classical consistency result and shows that weak dependence does not change the limiting geometry of nearest-neighbor radii. We then develop a non-asymptotic theory under geometric mixing. Under a local lower-mass condition of order s, we establish a tail bound for $R_{n,k}(x)$ and derive the sharp moment upper bound $\mathrm{E}[R_{n,k}(x)^p] \lesssim (k/n)^{p/s}$ in the standard regime $k \gtrsim \log n$. Under a matching local upper-mass condition of order s, we also establish the matching lower bound and hence obtain the exact order $\mathrm{E}[R_{n,k}(x)^p] \asymp (k/n)^{p/s}$. Thus, geometric dependence changes the concentration mechanism, but not the fundamental exponent.

A key feature of our theory is that it is formulated in terms of a local exponent $s > 0$, rather than the ambient dimension d. This point is especially relevant for modern machine learning, where high-dimensional data often concentrate near lower-dimensional latent structures [18, 3]. Our results therefore support the common regime in which the effective geometry around the query point is lower-dimensional than the ambient space. In this sense, our results are not only about dependence, but also about how dependence interacts with local geometric complexity.

Our contributions. The main contributions are as follows.

1. We establish *distribution-free almost sure convergence* for the k-NN radius under polynomial α-mixing. This extends the classical i.i.d. consistency result to dependent triangular arrays.

2. Under geometric α-mixing and a local lower-mass condition of order s, we establish a *non-asymptotic tail bound* for the k-NN radius and derive the *sharp moment upper bound* $\mathrm{E}[R_{n,k}(x)^p] \lesssim (k/n)^{p/s}$ for all $p > 0$ in the standard dependent regime $k \gtrsim \log n$.

3. Under a local upper-mass condition, we establish a *matching lower bound* $\mathrm{E}[R_{n,k}(x)^p] \gtrsim (k/n)^{p/s}$, and hence obtain the *exact-order characterization* $\mathrm{E}[R_{n,k}(x)^p] \asymp (k/n)^{p/s}$.

4. Our theory is stated in terms of a *local geometric exponent* s, which yields an intrinsic-dimension interpretation and is relevant for modern high-dimensional learning problems.

We believe the main message is simple and useful: under weak dependence, nearest-neighbor geometry remains essentially the same as in the independent case at the level of its canonical scale. This provides theoretical support for transferring geometric intuition and local-method design principles from the i.i.d. world to substantially broader dependent settings.

1.1 Related work

Applications of nearest-neighbor methods. In recent years, nearest-neighbor methods have re-emerged in modern machine learning as a flexible mechanism for retrieval, adaptation, and locality in learned representation spaces. Examples include self-supervised representation learning [4, 46], tabular deep learning [21], vector search and approximate nearest-neighbor infrastructure [12, 62, 30, 58], and robust retrieval and vision [59, 52]. Beyond core machine learning, nearest-neighbor structure also plays an important role in causal inference through matching estimators [1, 2, 39, 43, 10], in rank-based and graph-based measures of dependence [11, 5, 24, 40, 41, 42, 55, 25], in information-theoretic functional estimation via k-NN entropy and mutual-information estimators [34, 36], and in multivariate nonparametric testing based on geometric graphs [19, 28]. This breadth of applications makes the geometric behavior of nearest-neighbor neighborhoods a foundational issue.

Theory of nearest-neighbor methods under dependence. Despite the wide range of applications, previous literature on the theory under dependence has mostly analyzed estimators built from nearest-neighbor neighborhoods, rather than the geometry of the neighborhoods themselves. Examples include nearest-neighbor classification with dependent training sequences [29] and rates for nearest-neighbor estimation under more general non-i.i.d. sampling schemes [37]. More broadly, the literature on learning from dependent observations has developed both positive and negative results for nonparametric prediction under mixing, ergodic, and other non-i.i.d. assumptions [63, 53, 50]. By contrast, our focus is on the radius itself as a primary object. This shift in perspective is useful because the nearest-neighbor radius serves as the random local scale in many downstream methods, so controlling it directly provides a geometric ingredient for analyzing local learning methods. Related work gives oracle inequalities for regularized empirical risk minimization under stationary α-mixing [26], concentration for Banach-valued sums and kernel learning under τ-mixing [8], applications of U-statistic concentration for uniformly ergodic Markov chains [17], and generalization bounds for deep networks with τ-mixing samples [44]. These works analyze sums, risks, or testing statistics; our focus is the inversion of dependent small-ball counts to characterize the nearest-neighbor radius.

2 Setup

We study the geometry of nearest neighbors under dependent sampling. Let $[X_{n,i} : 1 \le i \le n,\ n \ge 1]$ be a triangular array of $\mathbb{R}^d$-valued random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathrm{P})$, where the ambient dimension $d \ge 1$ is fixed. For each row n, the observations $X_{n,1}, \dots, X_{n,n}$ could be independent, but we assume that they share a common marginal law μ on $\mathbb{R}^d$. Thus, for every $n \ge 1$ and every $i \in \{1, \dots, n\}$, $X_{n,i} \sim \mu$.

This formulation distinguishes two roles in our analysis. The common marginal distribution μ determines the local geometry of the marginal distribution, while the dependence structure controls how much effective information is contained in a sample of size n. Allowing the dependence structure to vary with n leads naturally to a triangular-array framework, which is flexible enough to cover the usual stationary-sequence setting as a special case and at the same time convenient for stating non-asymptotic results uniformly in n.

Triangular-array strong mixing. For integers $1 \le k \le m \le n$, let ${}^n\mathcal{F}_k^m := \sigma(X_{n,i} : k \le i \le m)$ denote the σ-field generated by the segment $(X_{n,k}, \dots, X_{n,m})$ in the nth row. Following the standard strong-mixing framework, define the mixing coefficients of the triangular array by

$$\alpha_m := \sup_{n \ge m+1} \sup_{1 \le k \le n-m} \sup \Big\{ \big| \mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B) \big| :\ A \in {}^n\mathcal{F}_1^k,\ B \in {}^n\mathcal{F}_{k+m}^n \Big\}, \qquad m \ge 1.$$

The array is said to be *strong mixing* (or α*-mixing*) if $\alpha_m \to 0$ as $m \to \infty$.

The strong-mixing framework is classical in probability and time-series analysis [9]. This definition requires a uniform decay of dependence across all rows of the array. Intuitively, if two collections of variables are separated by at least m indices within the same row, then their dependence becomes weak when m is large. This is the dependence notion used throughout the paper.

The triangular-array formulation is more general than working with a single stationary process $[X_i]_{i \ge 1}$. It allows the dependence structure to vary with n, while keeping the marginal distribution fixed. We state finite-sample results uniformly in n, which also covers the usual time-series setting

as a special case by taking $X_{n,i} = X_i$ for $1 \le i \le n$. Therefore, readers who prefer the standard sequential viewpoint may regard our framework as a generalization.

Nearest-neighbor radii. Fix a query point $x \in \mathbb{R}^d$. For each $n \ge 1$ and $k \in \{1, \dots, n\}$, let $\mathcal{X}_{(k)}^n(x) \in \{X_{n,1}, \dots, X_{n,n}\}$ denote the kth nearest neighbor of x among the sample points in the nth row, with ties broken arbitrarily. We introduce the k-NN radius:

$$R_{n,k}(x) := \|\mathcal{X}_{(k)}^n(x) - x\|,$$

together with the local counting process: $N_n(x, r) := \sum_{i=1}^n \mathbf{1}\{\|X_{n,i} - x\| \le r\}$. These two quantities are equivalent through the relationship $R_{n,k}(x) > r \iff N_n(x, r) < k$. This observation is the starting point for our entire analysis: upper bounds on the radius are equivalent to lower bounds for the number of sample points falling in $B(x, r)$. Under independence, $N_n(x, r)$ is binomial with parameters n and $\mu(B(x, r))$, so its behavior can be analyzed using classical binomial tail bounds. Under dependence, however, this analysis is not applicable since the indicators $\mathbf{1}\{\|X_{n,i} - x\| \le r\}$ are dependent, and hence the local counting process is not binomial. Controlling $R_{n,k}(x)$ therefore requires concentration arguments for dependent sequences, which is one of the main challenges.

Support and target limit. The asymptotic behavior of $R_{n,k}(x)$ depends on the position of the query point relative to the support of μ. Define $\mathrm{supp}(\mu) := \{x \in \mathbb{R}^d : \mu(B(x, r)) > 0 \text{ for all } r > 0\}$, where $B(x, r)$ denotes the closed Euclidean ball centered at x with radius r. We define the distance from x to the support of μ by $r_x := \inf_{y \in \mathrm{supp}(\mu)} \|y - x\|$. When $x \in \mathrm{supp}(\mu)$, one has $r_x = 0$; otherwise r_x is the deterministic limit that the k-NN radius converges to under mild assumptions.

Table 2 in Appendix A summarizes the notation used in the results and proofs.

3 Main results

We now present the main results for the k-NN radius under dependent sampling. As discussed in Section 2, the behavior of $R_{n,k}(x) = \|\mathcal{X}_{(k)}^n(x) - x\|$ depends on both the dependence structure of the sample and the local geometry of the marginal law around the query point x. The almost sure limit is determined by the position of x relative to the support of μ, whereas the convergence rate depends on how the mass of μ accumulates near x.

3.1 Distribution-free almost sure convergence

Our first theorem shows the almost sure convergence of the k-NN radius to r_x under weak dependence.

Theorem 1. *Assume that the triangular array $[X_{n,i}]_{1 \le i \le n,\ n \ge 1}$ is α-mixing with coefficients $[\alpha_m]_{m \ge 1}$ satisfying $\alpha_m \lesssim m^{-\gamma}$ for some $\gamma > 1$. If $k/n \to 0$, then $R_{n,k}(x) \xrightarrow{\text{a.s.}} r_x$. In particular, if $x \in \mathrm{supp}(\mu)$ and $k/n \to 0$, then $R_{n,k}(x) \xrightarrow{\text{a.s.}} 0$.*

Theorem 1 may be viewed as the dependent-sampling analogue of the classical i.i.d. consistency result [7, Lemma 2.2]. It shows that weak dependence, quantified here through a polynomial mixing condition, does not change the limiting geometric behavior of nearest-neighbor radii. Importantly, the result remains distribution-free: no density, smoothness, or local mass condition is required. The only assumption is that dependence decays polynomially fast along each row of the triangular array.

3.2 From almost sure convergence to moment bounds

We next turn to non-asymptotic moment results. For these stronger results, one must quantify how much probability mass lies near the query point x.

Rather than working with the ambient dimension d, we state the results using a local exponent $s > 0$. This formulation is more flexible and captures the possibility that the effective local dimension near x is smaller than the ambient dimension. The standard Euclidean case is recovered by setting $s = d$.

For the non-asymptotic results, we impose the stronger assumption that the mixing coefficients decay exponentially fast.

Assumption 1 (Geometric α-mixing)**.** *There exists $\gamma > 0$ such that $\alpha_m \le e^{-\gamma m}$ for all $m \ge 1$.*

Assumption 1 strengthens the assumption in Theorem 1 from polynomial to exponential decorrelation. This stronger assumption enables non-asymptotic concentration via blocking arguments. The parameter γ quantifies the strength of dependence: larger γ corresponds to faster decorrelation, and in the limit of vanishing dependence, one recovers behavior close to the i.i.d. case.

Geometric mixing is a standard assumption in time-series theory and is satisfied by many classical models under mild regularity conditions [16]. In particular, stationary ARMA processes are known to be strongly mixing with exponential rate under standard stability and innovation-regularity conditions [51], which makes geometric mixing a natural benchmark for studying dependent nearest-neighbor geometry.

Remark 1 (Other mixing conditions)**.** *We use α-mixing because it directly controls measurable ball events and is weaker than the β- and φ-mixing conditions. The radius bounds extend immediately to β-mixing and φ-mixing, at the same polynomial rate as in Theorem 1 and the same geometric rate as in Assumption 1. They do not extend to $\mathcal{C}$-mixing by the arguments of this paper. For any two σ-fields, $2\alpha(\mathcal{A},\mathcal{B}) \le \beta(\mathcal{A},\mathcal{B}) \le \varphi(\mathcal{A},\mathcal{B})$ [9, eq. (1.11)], and the same comparison holds for the coefficients of each row, so $\alpha_m \le \beta_m/2 \le \varphi_m/2$. A triangular array whose β- or φ-coefficients satisfy either decay condition is therefore α-mixing at that same rate, and every result applies with no further argument. The converse implications fail in general [9, Section 2.1].*

The obstruction for $\mathcal{C}$-mixing is the form of the statistics in the proofs. $\mathcal{C}$-mixing controls correlations against a prescribed class of regular functions, such as Lipschitz or Hölder functions [47, 27]. Ball indicators are discontinuous on the sphere and do not have a finite Lipschitz or Hölder $\mathcal{C}$-norm, so those inequalities do not apply directly to $N_n(x,r)$; a separate approximation argument would be needed.

Assumption 2 (Local lower mass of order s at x)**.** *There exist constants $c_- > 0$, $r_0 > 0$, and $s > 0$ such that $\mu(B(x,r)) \ge c_- r^s$ for all $0 < r \le r_0$.*

Assumption 2 guarantees that the ball $B(x,r)$ contains enough probability mass for its radius-r neighborhood to capture roughly nr^s sample points. This lower bound is the key ingredient for the upper bound of $R_{n,k}(x)$.

Assumption 3 (Local upper mass of order s at x)**.** *There exist constants $c_+ > 0$, $r_0 > 0$, and $s > 0$ such that $\mu(B(x,r)) \le c_+ r^s$ for all $0 < r \le r_0$.*

Assumption 3 matches Assumption 2 and prevents the distribution from placing too much mass near x, and is needed to establish a matching lower bound on the radius.

Assumption 4 (Compact support)**.** *The support of μ is contained in a compact set of diameter D.*

Assumption 4 is for technical simplicity rather than a geometric limitation. The local behavior of the k-NN radius is determined by Assumptions 2 and 3, which quantify how the mass of μ accumulates near the query point x. By contrast, compact support is used to control the far tail of $R_{n,k}(x)$ when converting non-asymptotic tail bounds into moment bounds. Proposition B.6 replaces compact support by $\mathrm{E}\|X_{n,1} - x\|^q < \infty$ for some $q > p$, covering the Gaussian design of Experiment 2.

Interpretation of the local exponent s. The exponent s is the *local geometric dimension* around the query point x. Assumptions 2 and 3 say that, for sufficiently small radii, $\mu(B(x,r)) \asymp r^s$. Thus, the amount of probability mass captured by a small neighborhood of x grows like the volume of an s-dimensional ball. In the standard Euclidean case with a positive density in $\mathbb{R}^d$, one has $s = d$. More generally, s may be strictly smaller than d, e.g., when the distribution is supported on a lower-dimensional manifold, subspace, or other latent geometric structure. Fixed full-dimensional noise can instead give $s = d$ at sufficiently small radii. Our theory is therefore local and intrinsic: the scaling of the k-NN radius is determined by the intrinsic dimension of the distribution near x, rather than by the ambient Euclidean dimension.

Throughout the sharp upper-bound results, we work in the regime $K_0 \log n \le k \le \kappa n$, for sufficiently large $K_0 > 0$ and $0 < \kappa < c_- r_0^s/8$. The threshold K_0 may depend on the desired tail exponent or moment order. The upper bound ensures that the reference scale $r_*(n,k)$ lies within the local regime $r \le r_0$, where Assumption 2 applies. The lower bound $k \gtrsim \log n$ is sufficient for our blocking proof: it arises because the proof separates the sample into blocks of length $\asymp \log n$, as is standard in Bernstein-type concentration arguments for geometric mixing sequences [38, 49]. We do not establish that this threshold is necessary. In contrast, the lower bound on the radius does not require $k \gtrsim \log n$.

3.3 Non-asymptotic upper bound

We first establish a tail bound on $R_{n,k}(x)$.

Theorem 2 (Non-asymptotic tail bound)**.** *Assume Assumptions 1 and 2. Fix $A > 0$ and define $b_{n,A} := \lceil \frac{A+1}{\gamma} \log n \rceil$, $r_*(n,k) := (\frac{8k}{c_- n})^{1/s}$. There exist constants $c_A, C_A, K_A > 0$, depending only on A and γ, such that, whenever $K_A \log n \le k \le \kappa n$, for every integer $j \ge 0$ satisfying $2^j r_*(n,k) \le r_0$, we have*

$$\mathrm{P}\big(R_{n,k}(x) > 2^j r_*(n,k)\big) \le 4 \exp\Big(- c_A \, 2^{js} \, \frac{k}{\log n} \Big) + C_A n^{-A},$$

for all sufficiently large n.

Theorem 2 shows that the k-NN radius is concentrated around the canonical geometric scale $(k/n)^{1/s}$. In particular, Theorem 2 separates the roles of geometry and dependence: the exponent $1/s$ is determined by the local mass behavior of μ near x, while the dependence structure enters only through the additional $\log n$ factor in the concentration exponent, which is the price of the blocking argument under geometric mixing. Thus, although dependence weakens concentration relative to the i.i.d. case, it does not change the underlying nearest-neighbor scale. For $t \ge 1$, Theorem 2 gives $R_{n,k}(x) \le C((k + t \log n)/n)^{1/s}$ with probability at least $1 - 4e^{-c_A t} - C_A n^{-A}$, provided this upper radius is at most r_0 (after adjusting constants). Choosing $A > p/s$ makes the remainder small enough to obtain the moment bound for every $p > 0$.

Integrating the tail bound yields the sharp moment upper bound.

Theorem 3 (Sharp moment upper bound)**.** *Assume Assumptions 1, 2, and 4. Let $p > 0$. Then there exist constants $K_0, C_{p,s} > 0$, depending only on $(p, s, \gamma, c_-, r_0, D)$, such that, for all sufficiently large n, whenever $K_0 \log n \le k \le \kappa n$, we have $\mathrm{E}\big[R_{n,k}(x)^p\big] \le C_{p,s}(k/n)^{p/s}$.*

If μ admits a density, then by the Lebesgue differentiation theorem, Assumption 2 holds with $s = d$ for μ-almost every interior point at which the density is positive. Consequently, Theorem 3 applies for μ-almost all such x. Theorem 3 is applicable in downstream analyses, since many local learning bounds depend on moments of the random neighborhood. We discuss this further in Section 5.

3.4 Matching non-asymptotic lower bound

To show that the exponent p/s is optimal, we next establish a matching lower bound. Unlike the upper bound, the lower bound is purely local and does not require any dependence condition.

Theorem 4 (Matching lower bound)**.** *Assume Assumption 3. Let $p > 0$. Then whenever $(\frac{k}{2c_+ n})^{1/s} \le r_0$, we have $\mathrm{E}\big[R_{n,k}(x)^p\big] \ge [k/(2c_+ n)]^{p/s}/2$.*

Theorem 4 is conceptually simple but important: if a ball of radius r contains at most order r^s probability mass, then one cannot expect the kth nearest neighbor to lie much closer than $(k/n)^{1/s}$. Combining the upper and lower bounds yields the following exact-order characterization.

Theorem 5 (Exact-order characterization)**.** *Assume Assumptions 1, 2, 3, and 4, with the same exponent $s > 0$. Then for every $p > 0$ there exist a threshold K_0 and constants $0 < c_{p,s} \le C_{p,s} < \infty$, depending only on $(p, s, \gamma, c_-, c_+, r_0, D)$, such that, for all sufficiently large n and all $K_0 \log n \le k \le \kappa n$, we have $c_{p,s}(k/n)^{p/s} \le \mathrm{E}\big[R_{n,k}(x)^p\big] \le C_{p,s}(k/n)^{p/s}$.*

Theorem 5 shows that the nearest-neighbor radii scale remains unchanged under geometric mixing. The dependence changes constants and concentration behavior, but not the fundamental exponent.

4 Experiments

4.1 Moment convergence rate

We test whether the predicted moment rate remains visible across different geometrically α-mixing processes and dependence strengths. The implementation compares the i.i.d. benchmark with three dependent sequence models: a linear Gaussian state-space model, a hidden Markov model, and a

stationary Gaussian-process sequence with a radial basis function (RBF)-like kernel generated by fast Fourier transform (FFT). All observations are marginally transformed to the cube $[0,1]^d$.

We sample 1000 points uniformly from the interior of the cube and report the mean p-th moment of the k-NN radius. For each $d \in \{1,3,5\}$, we choose p such that $p/d \in \{1,2,3,4,5\}$. We use exponentially growing, dimension-dependent sample sizes $n(d) = 100 \cdot 2^d \cdot m$ for $m \in \{1,2,4,8,16,32\}$. The number of NNs is chosen as $k = \lfloor n^\beta \rceil$ with β on a 20-point grid in $[0.1, 0.9]$ for slope fitting, subject to $k/n \leq 0.01$. The slope is the coefficient b from $\log M = a + b \log(k/n)$, with theoretical slope p/d for the i.i.d. and α-mixing setting.

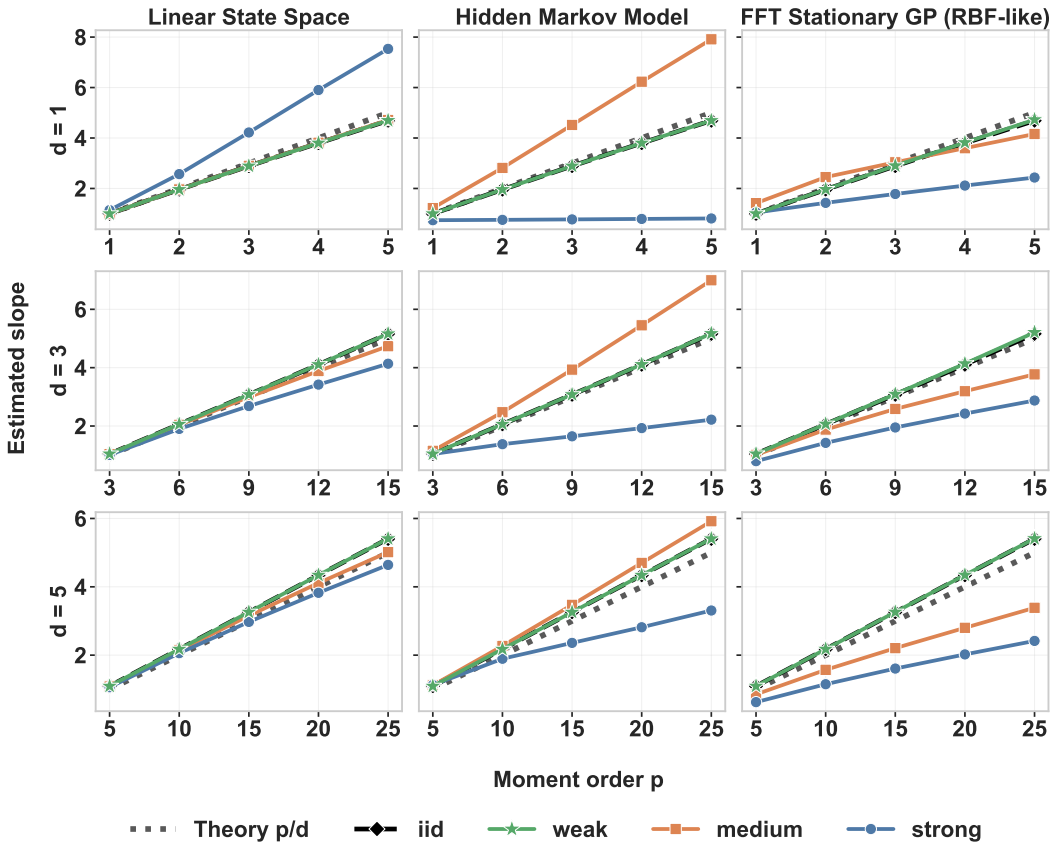


Figure 1: Slope summary for Experiment 1: rows correspond to $d = 1, 3, 5$, columns correspond to the three dependent model classes, and the shared legend compares the theoretical line p/d, the i.i.d. benchmark, and weak/medium/strong dependence.

These generators satisfy Assumption 1; explicit rates are given in Propositions B.1, B.2, and B.3, and the local-mass exponents are checked in Appendix B.4. The departures from p/d in Figure 1 occur for the strong-dependence settings, whose measured effective sample size at $n = 6400$ and $d = 1$ is at most 90, while every configuration with $n_{\text{eff}} \gtrsim 2500$ matches the i.i.d. slope error (Figure 7 and Table 4).

Across the three data-generating processes (linear state-space, hidden Markov, and FFT Gaussian process), we impose weak/medium/strong dependence via model-specific persistence parameters. Figure 1 shows the expected pattern: weak dependence stays close to the i.i.d. baseline and the theoretical rate p/d, while medium and strong dependence diverge.

4.2 Entropy estimation

This experiment examines the intrinsic-dimension message of the nearest-neighbor theory in a representation-learning setting. We state our results in terms of a local intrinsic dimension s. A Kozachenko–Leonenko entropy estimate is built from the NN radii and the corresponding s-

dimensional volume correction. We use this estimator to illustrate the effect of a representation that preserves the local s-dimensional geometry; consistency of the entropy estimator is not implied by our fixed-query radius results. We consider the setup in which observations have ambient dimension d, but the true data-generating process has intrinsic dimension s.

The experiment generates an s-dimensional Gaussian AR(1) latent process, whose stationary marginal differential entropy is available in closed form as $H_s = (s/2)\log(2\pi e)$, and maps it into 20-dimensional observed data through a linear augmentation. Figure 2 reports the results for $s \in \{1, 3, 5\}$ under three temporal-dependence regimes for the sample sequence: i.i.d. ($\rho = 0$), weak AR(1) dependence ($\rho = 0.3$), and strong AR(1) dependence ($\rho = 0.6$), using 1000 replications and number of NNs $k = \max\{2, \lceil n^{0.1} \rceil\}$. The oracle estimator on the latent s-dimensional data and the estimator based on the top s coordinates of a 5-component principal component analysis (PCA) representation both move toward the true entropy as n grows. In this noiseless design the embedding is an isometry, and the top s PCA coordinates preserve pairwise distances once the centered sample spans the signal subspace. Thus the two entropy estimates agree up to numerical precision; this is not a guarantee for arbitrary learned representations. The schedule $k = \lceil n^{0.1} \rceil$ eventually exceeds any fixed multiple of $\log n$, although finite sample sizes need not meet the theorem's threshold. Implementation details are given in Appendix B.2.

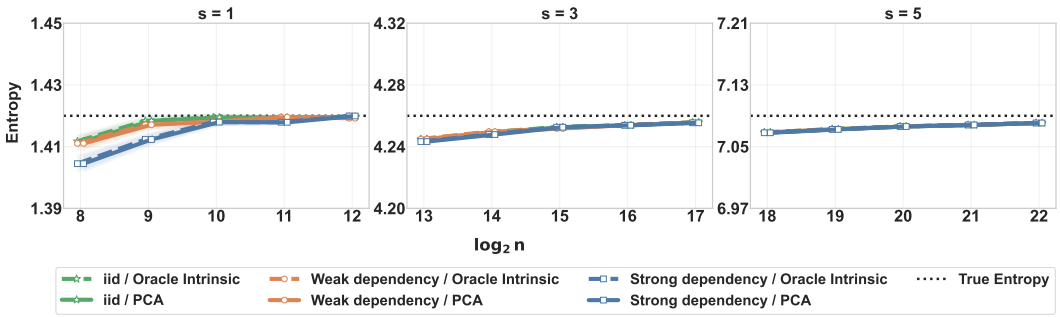


Figure 2: Experiment 2 entropy estimation. The dotted horizontal line is the true entropy. Colors encode the dependence level and markers distinguish the oracle intrinsic estimate from the PCA estimate.

The same design satisfies the assumptions with intrinsic exponent s rather than ambient dimension 20. Remark 4 shows that the latent AR(1) is geometrically α-mixing, that the fixed orthonormal embedding preserves both the mixing rate and the local ball mass of order s, and that the Gaussian marginal is covered by the moment bound in Proposition B.6. Figure 9 shows the empirical ball mass in $\mathbb{R}^{20}$ lying on the exact s-dimensional curves, with fitted slopes about 2.84 and 4.81–4.85 for targets $s = 3$ and $s = 5$.

4.3 Real-data evaluation

In this section, we complement the synthetic experiments with real-world time-series benchmarks. We compare two baselines: a raw k-nearest-neighbors (Raw-kNN) method operating directly in standardized observation space, and a PCA k-nearest-neighbors (PCA-kNN) method that first projects each input onto a lower-dimensional representation using PCA before applying nearest-neighbor prediction. We evaluate Raw-kNN and PCA-kNN on three real-data tasks: long-horizon forecasting, short-horizon forecasting, and classification. The datasets are from the Time Series Pile [22], a large public time-series benchmark collection that includes the Informer long-horizon forecasting datasets [64], the Monash Time Series Forecasting Archive [20], and the UCR Time Series Archive [14]. We follow the benchmark setup and compare our results against MOMENT, a family of pre-trained transformer models for general-purpose time-series analysis [22]. Forecasting predicts future values from past windows, while classification assigns a label to an entire series. In both nearest-neighbor models, the input features are standardized before nearest-neighbor prediction. We tune the number of neighbors $k \in \{1, 3, 5, 7, 9, 15, 21, 31\}$, the neighbor weighting scheme (uniform/distance), and the distance metric (Manhattan/Euclidean). For PCA-kNN, we additionally tune the embedding

dimension over $\{8, 16, 32, 48, 64\}$. Dataset definitions, metrics, preprocessing, train/validation/test splits, and tuning protocols for all three tasks are provided in Appendix B.5.

Table 1 summarizes average performance across datasets for long-horizon forecasting, short-horizon forecasting, and classification. Figure 3 reports long-horizon forecasting mean squared error (MSE) separately for each dataset and forecast horizon. Figure 4 provides complementary views of short-horizon forecasting and classification: its left panel shows short-horizon forecasting symmetric mean absolute percentage error (sMAPE) under both the within-group and source-to-target transfer settings, while its right panel shows the distribution of classification accuracy with MOMENT summary statistics included as reference lines. Together, these results are supportive in two main ways. First, Raw-kNN and PCA-kNN remain informative across all three task families. In long-horizon forecasting, the two methods achieve average MSE 0.628 and 0.627, respectively, compared with 0.617 for MOMENT. In short-horizon forecasting, both methods produce nontrivial predictions in both the within-group and source-to-target transfer settings, although errors increase substantially in the transfer setting. In classification, both methods achieve mean accuracy 0.748, compared with 0.794 for MOMENT. These findings indicate that nearest-neighbor structure remains informative on real dependent data. Second, Raw-kNN and PCA-kNN achieve similar performance across tasks, consistent with dimension reduction preserving information useful for nearest-neighbor prediction. Their relative advantage varies across datasets, but neither method consistently dominates the other.

Although MOMENT generally outperforms the Raw-kNN and PCA-kNN baselines, this does not contradict the theory. Both nearest-neighbor methods remain effective on real dependent data, and PCA-kNN performs comparably to Raw-kNN even after reducing the input dimension. This suggests that dimension reduction can preserve the local structure needed for neighborhood-based prediction. Appendix B.6 directly examines radius scaling and local mass: fitted radius slopes are positive across the reported datasets, with 95% query-block-bootstrap intervals bounded away from zero. These finite-range diagnostics do not establish the exact exponent $1/s$ or verify α-mixing; the intervals condition on the reference data and fitted representations.

Table 1: Comparison of Raw-kNN, PCA-kNN, and MOMENT on real-data tasks under a common benchmark setup. Performance is averaged over datasets within each task; MOMENT results are taken from [22].

Task	Metric	Raw-kNN	PCA-kNN	MOMENT
Long-horizon forecasting	MSE	0.628	0.627	0.617
Classification	Accuracy	0.748	0.748	0.794
Short-horizon forecasting (within-group)	sMAPE	16.199	16.488	–
Short-horizon forecasting (transfer)	sMAPE	27.210	27.369	14.525

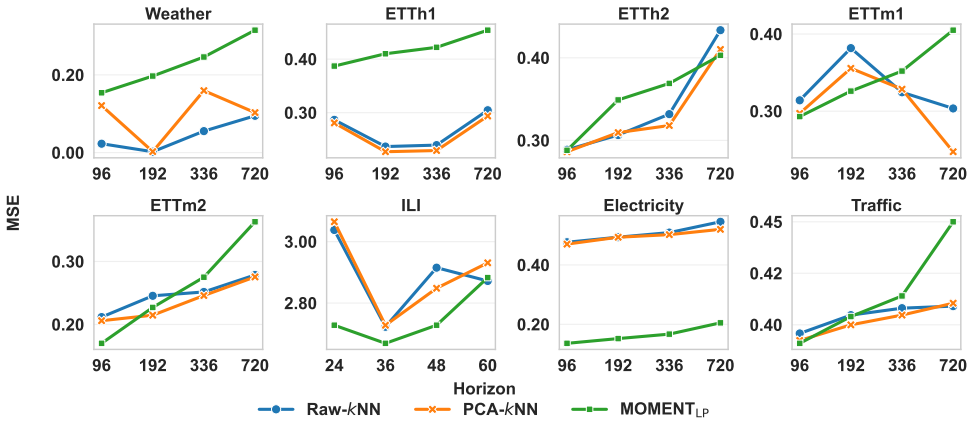


Figure 3: Comparison of long-horizon forecasting using test MSE across datasets and forecast horizons. It compares Raw-kNN and PCA-kNN with MOMENT across benchmark settings.

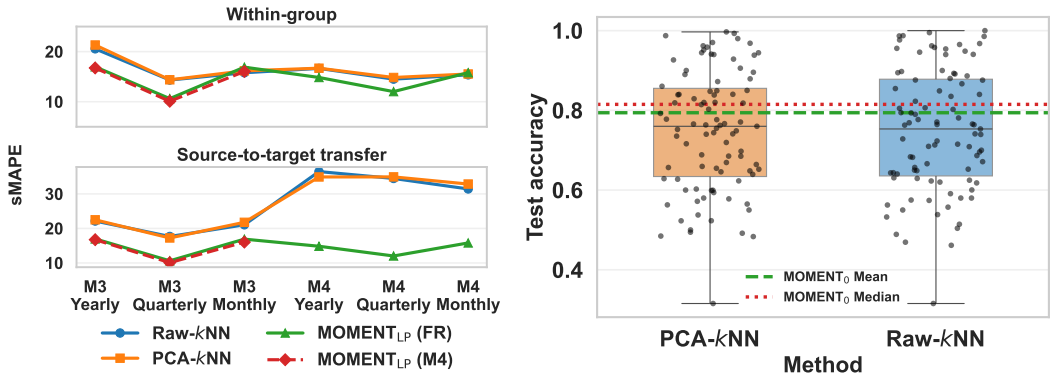


Figure 4: Comparisons on short-horizon forecasting and classification. The left panel compares forecasting under within-group and source-to-target transfer settings. The right panel shows classification accuracy across UCR datasets, with MOMENT summary statistics shown as reference lines.

5 Discussions and limitations

Constants versus exponents. One main message of the theory is that dependence and local geometry play qualitatively different roles. The local exponent s determines the fundamental scale $(k/n)^{1/s}$, whereas the dependence structure determines constants and concentration quality. This distinction is important both conceptually and empirically: it suggests that weak dependence should affect finite-sample behavior without changing the asymptotic behavior.

On uniform convergence. We study nearest-neighbor radii at a fixed query point x. Uniform control requires a lower-mass condition with common (c_-, s, r_0) over a deterministic query set Q, together with control of its covering numbers. For an ε-net of Q with M centers in Q, a union bound and the 1-Lipschitz property of $x \mapsto R_{n,k}(x)$ give

$$\mathrm{P}\left(\sup_{x\in Q} R_{n,k}(x) > 2^j r_* + \varepsilon\right) \le M\left[4e^{-c_A 2^{js} k/\log n} + C_A n^{-A}\right], \qquad 2^j r_* \le r_0.$$

Thus covering entropy enters the concentration bound, and A must also control the accumulated remainder. The same union bound applies to a fixed finite set of queries. Sample-dependent queries require a uniform event covering their possible locations; they cannot simply be substituted into a fixed-query bound. A general uniform theory for downstream estimators requires further assumptions and is left to future work [63, 31, 23].

Time-series representations. The synthetic proofs cover Gaussian state-space, hidden Markov, and GP–RBF models (Appendix B.3); stationary causal ARMA models are also covered under the regularity conditions of Mokkadem [51]. A fixed measurable frame-level encoder cannot increase α-mixing coefficients. If it is bi-Lipschitz on the support, it also preserves the local mass exponent, with changed constants. Neither conclusion automatically extends to an encoder fitted on the same dependent sample, and autocorrelation decay alone does not verify strong mixing.

Implications for learning methods. More broadly, we provide a geometric control principle for a wide class of learning methods under dependent sampling. Since the k-NN radius determines the effective neighborhood size around a query point, our results imply that, under the stated assumptions, dependent observations behave as if the relevant scale were still $(k/n)^{1/s}$ up to constants and a dependence-induced concentration penalty. In particular, whenever the performance of a method is driven by the size of adaptive neighborhoods—as in k-NN regression, local averaging, adaptive bandwidth selection, graph construction, or neighborhood-based manifold methods—our results can be used as an ingredient to analyze their performance on dependent data. Detailed development of these implications for nonparametric and graph learning algorithms is left for future work.

Acknowledgments

We thank Fang Han for providing part of the motivation for this work and for insightful comments and discussions. We also thank the anonymous reviewers, Area Chairs, and Senior Area Chairs for their valuable feedback and suggestions.

Funding and competing interests

The authors received no external funding or third-party support for this work. The authors declare no competing interests.

References

[1] Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

[2] Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11, 2011.

[3] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Dmitrii Avdiukhin, Vaggos Chatziafratis, Orr Fischer, and Grigory Yaroslavtsev. Embedding dimension of contrastive learning and k-nearest neighbors. *Advances in Neural Information Processing Systems*, 37:41359–41393, 2024.

[5] Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *The Annals of Statistics*, 49(6):3070–3102, 2021.

[6] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[7] Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015.

[8] Gilles Blanchard and Oleksandr Zadorozhnyi. Concentration of weakly dependent Banach-valued sums and applications to statistical learning methods. *Bernoulli*, 25(4B):3421–3458, 2019.

[9] Richard C Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

[10] Matias D Cattaneo, Fang Han, and Zhexiao Lin. On rosenbaum's rank-based matching estimator. *Biometrika*, 112(1):asae062, 2025.

[11] Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021.

[12] Qi Chen, Bing Zhao, Haidong Wang, Mingqin Li, Chuanjie Liu, Zengzhong Li, Mao Yang, and Jingdong Wang. Spann: Highly-efficient billion-scale approximate nearest neighborhood search. *Advances in Neural Information Processing Systems*, 34:5199–5212, 2021.

[13] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[14] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

[15] Yu A Davydov. Convergence of distributions generated by stationary stochastic processes. *Theory of Probability & Its Applications*, 13(4):691–696, 1968.

[16] Paul Doukhan. *Mixing: Properties and Examples*, volume 85. Springer Science & Business Media, 2012.

[17] Quentin Duchemin, Yohann De Castro, and Claire Lacour. Three rates of convergence or separation via U-statistics in a dependent framework. *Journal of Machine Learning Research*, 23(201):1–59, 2022.

[18] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[19] Jerome H Friedman and Lawrence C Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979.

[20] Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643*, 2021.

[21] Yury Gorishniy, Ivan Rubachev, Nikolay Kartashev, Daniil Shlenskii, Akim Kotelnikov, and Artem Babenko. Tabr: Tabular deep learning meets nearest neighbors in 2023. *arXiv preprint arXiv:2307.14338*, 2023.

[22] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024.

[23] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

[24] Fang Han. On extensions of rank correlation coefficients to multivariate spaces. *Bernoulli News*, 28(2):7–11, 2021.

[25] Fang Han and Zhihan Huang. Azadkia–chatterjee's correlation coefficient adapts to manifold data. *The Annals of Applied Probability*, 34(6):5172–5210, 2024.

[26] Hanyuan Hang and Ingo Steinwart. Fast learning from α-mixing observations. *Journal of Multivariate Analysis*, 127:184–199, 2014.

[27] Hanyuan Hang and Ingo Steinwart. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics*, 45(2):708–743, 2017.

[28] Norbert Henze and Mathew D Penrose. On the multivariate runs test. *The Annals of Statistics*, 27(1):290–298, 1999.

[29] M Holst and A Irle. Nearest neighbor classification with dependent training sequences. *The Annals of Statistics*, 29(5):1424–1442, 2001.

[30] Elias Jääsaari, Ville Hyvönen, and Teemu Roos. Lorann: Low-rank matrix factorization for approximate nearest neighbor search. *Advances in Neural Information Processing Systems*, 37:102121–102153, 2024.

[31] Heinrich Jiang. Non-asymptotic uniform rates of consistency for k-nn regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3999–4006, 2019.

[32] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. In *International Conference on Learning Representations*, 2021.

[33] Andrei Nikolaevich Kolmogorov and Yu A Rozanov. On strong mixing conditions for stationary gaussian processes. *Theory of Probability & Its Applications*, 5(2):204–208, 1960.

[34] LF Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

[35] Samory Kpotufe. k-nn regression adapts to local intrinsic dimension. *Advances in Neural Information Processing Systems*, 24, 2011.

[36] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

[37] Sanjeev R Kulkarni and Steven E Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.

[38] Eckhard Liebscher. Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stochastic Processes and Their Applications*, 65(1):69–80, 1996.

[39] Zhexiao Lin, Peng Ding, and Fang Han. Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica*, 91(6):2187–2217, 2023.

[40] Zhexiao Lin and Fang Han. Limit theorems of chatterjee's rank correlation. *arXiv preprint arXiv:2204.08031*, 2022.

[41] Zhexiao Lin and Fang Han. On boosting the power of chatterjee's rank correlation. *Biometrika*, 110(2):283–299, 2023.

[42] Zhexiao Lin and Fang Han. On the failure of the bootstrap for chatterjee's rank correlation. *Biometrika*, 111(3):1063–1070, 2024.

[43] Zhexiao Lin and Fang Han. On regression-adjusted imputation estimators of average treatment effects. *Journal of Econometrics*, 251:106080, 2025.

[44] Liyuan Liu, Yaohui Chen, Weifu Li, Yingjie Wang, Bin Gu, Feng Zheng, and Hong Chen. Generalization bounds of deep neural networks with τ-mixing samples. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):14596–14610, 2025.

[45] Don O Loftsgaarden and Charles P Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

[46] Xianzhong Long, Chen Peng, and Yun Li. Mnn: Mixed nearest-neighbors for self-supervised learning. *Pattern Recognition*, 158:110998, 2025.

[47] Véronique Maume-Deschamps. Exponential inequalities and functional estimations for weak dependent data; applications to dynamical systems. *Stochastics and Dynamics*, 6(4):535–560, 2006.

[48] Daniel J McDonald, Cosma Rohilla Shalizi, and Mark Schervish. Estimating beta-mixing coefficients via histograms. *Electronic Journal of Statistics*, 9:2855–2883, 2015.

[49] Florence Merlevède, Magda Peligrad, and Emmanuel Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, volume 5, pages 273–292. Institute of Mathematical Statistics, 2009.

[50] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11:661–686, 2010.

[51] Abdelkader Mokkadem. Mixing properties of arma processes. *Stochastic Processes and Their Applications*, 29(2):309–315, 1988.

[52] Ori Nizan and Ayellet Tal. k-nnn: nearest neighbors of neighbors for anomaly detection. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 1005–1014, 2024.

[53] Andrew B Nobel. Limits to classification and regression estimation from ergodic processes. *The Annals of Statistics*, 27(1):262–273, 1999.

[54] Emmanuel Rio. The functional law of the iterated logarithm for stationary strongly mixing sequences. *The Annals of Probability*, 23(3):1188–1203, 1995.

[55] Hongjian Shi, Mathias Drton, and Fang Han. On azadkia–chatterjee's conditional dependence coefficient. *Bernoulli*, 30(2):851–877, 2024.

[56] Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977.

[57] Stephen Tu, Roy Frostig, and Mahdi Soltanolkotabi. Learning from many trajectories. *Journal of Machine Learning Research*, 25(216):1–109, 2024.

[58] Mengzhao Wang, Xiaoliang Xu, Qiang Yue, and Yuxiang Wang. A comprehensive survey and experimental comparison of graph-based approximate nearest neighbor search. *arXiv preprint arXiv:2101.12631*, 2021.

[59] Yihan Wu, Hongyang Zhang, and Heng Huang. Retrievalguard: Provably robust 1-nearest neighbor image retrieval. In *International Conference on Machine Learning*, pages 24266–24279. PMLR, 2022.

[60] Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, and Yongdong Zhang. k nn prompting: Beyond-context learning with calibration-free nearest neighbor inference. *arXiv preprint arXiv:2303.13824*, 2023.

[61] Frank F Xu, Uri Alon, and Graham Neubig. Why do nearest neighbor language models work? In *International Conference on Machine Learning*, pages 38325–38341. PMLR, 2023.

[62] Ming Yang, Yuzheng Cai, and Weiguo Zheng. Cspg: crossing sparse proximity graphs for approximate nearest neighbor search. *Advances in Neural Information Processing Systems*, 37:103076–103100, 2024.

[63] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

[64] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[65] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

Table 2: Notation used in the results and proofs.

Symbol	Meaning
$X_{n,i}$, μ	Observations and their common marginal law
$R_{n,k}(x)$, $N_n(x,r)$	k-NN radius and number of observations in $B(x,r)$
r_x, p_r	Distance to $\mathrm{supp}(\mu)$ and ball mass $\mu(B(x,r))$
s, $c_\pm$, r_0	Local mass exponent, constants, and radius of validity
$Y_i(r)$, $Z_{n,i}$	Ball indicator and centered indicator, respectively
$b_{n,A}$, $W_u(r)$, V_u	Block length, kept-block count, and centered block count

A Proofs

The proofs use the equivalence $R_{n,k}(x) > r \iff N_n(x,r) < k$. For almost sure convergence, a Bernstein bound at a fixed radius gives summable probabilities. For the finite-sample upper bounds, we retain alternating blocks of length $O(\log n)$, control their mixing coefficients, and integrate the resulting tail bound. The matching lower bound uses only the common marginal law and Markov's inequality.

We begin with a Bernstein-type inequality for α-mixing sequences. The alternative inequality of Merlevède, Peligrad, and Rio [49, Theorem 2] involves a long-run covariance sum and a $(\log n)^2$ term. Lemma 1 is convenient here because the kept-block variance bound retains the small-ball probability and the block length makes the mixing remainder arbitrarily small. Different inequalities may alter constants and logarithmic thresholds, but the matching lower bound shows that the geometric exponent p/s cannot be improved under the stated local-mass assumptions.

Lemma 1 (Theorem 5 of [54]; Theorem 2.1 of [38])**.** *Let $[Z_i]_{i=1}^n$ be a centered sequence of random variables such that*

$$\mathrm{E}[Z_i] = 0 \qquad \textit{and} \qquad |Z_i| \le S(n) < \infty, \qquad i \in \{1,\dots,n\}.$$

Then, for any $n \ge 1$, any $m \in \{1,\dots,n\}$, and any $\epsilon > 4mS(n)$,

$$\mathrm{P}\Big(\Big|\sum_{i=1}^n Z_i\Big| > \epsilon\Big) \le 4\exp\left(-\frac{\epsilon^2}{64\frac{n}{m}\sigma_{n,m}^2 + \frac{8}{3}\,\epsilon m S(n)}\right) + 4\frac{n}{m}\alpha_m,$$

where

$$\sigma_{n,m}^2 := \sup_{0\le j\le n-1} \mathrm{E}\Big[\Big(\sum_{i=j+1}^{(j+m)\wedge n} Z_i\Big)^2\Big].$$

A.1 Proof of Theorem 1

Proof. Fix $x \in \mathbb{R}^d$. Since $\mathcal{X}_{(k)}^n(x) \in \mathrm{supp}(\mu)$ almost surely and $r_x = \inf\{\|y-x\| : y \in \mathrm{supp}(\mu)\}$, we have

$$R_{n,k}(x) \ge r_x \qquad \text{a.s. for every } n.$$

Thus, it remains to prove that

$$\limsup_{n\to\infty} R_{n,k}(x) \le r_x \qquad \text{a.s.}$$

Fix $\varepsilon > 0$, and define

$$p_x := \mu(B(x, r_x+\varepsilon)) = \mathrm{P}\big(\|X_{n,i} - x\| \le r_x + \varepsilon\big).$$

Since the interior of $B(x, r_x+\varepsilon)$ intersects $\mathrm{supp}(\mu)$ by definition of r_x, we have $p_x > 0$.

We observe that the event $R_{n,k}(x) > r_x + \varepsilon$ occurs if and only if fewer than k observations in the n-th row fall inside $B(x, r_x+\varepsilon)$. Then we have

$$\mathrm{P}\big(R_{n,k}(x) > r_x+\varepsilon\big) = \mathrm{P}\Big(\sum_{i=1}^n \mathbf{1}\{\|X_{n,i}-x\| \le r_x+\varepsilon\} < k\Big).$$

We set

$$Z_{n,i} := \mathbf{1}\{\|X_{n,i} - x\| \le r_x + \varepsilon\} - \mathrm{P}\big(\|X_{n,i} - x\| \le r_x + \varepsilon\big) = \mathbf{1}\{\|X_{n,i} - x\| \le r_x + \varepsilon\} - p_x.$$

Then $\mathrm{E}[Z_{n,i}] = 0$ and $|Z_{n,i}| \le 1$. Since $Z_{n,i}$ is a measurable function of $X_{n,i}$, the sequence $[Z_{n,i}]_{i=1}^n$ is again α-mixing with mixing coefficients no greater than those of $[X_{n,i}]_{i=1}^n$. Moreover, for any $m \in \{1, \dots, n\}$ and any $0 \le j \le n-1$, the covariance inequality [15, Corollary of Lemma 2.1] yields

$$\begin{aligned}
\mathrm{E}\Big[\Big(\sum_{i=j+1}^{(j+m)\wedge n} Z_{n,i}\Big)^2\Big] &= \mathrm{Var}\Big(\sum_{i=j+1}^{(j+m)\wedge n} Z_{n,i}\Big) \\
&\le \sum_{i=j+1}^{(j+m)\wedge n} \mathrm{Var}(Z_{n,i}) + 2 \sum_{i=j+1}^{(j+m)\wedge n} \sum_{k=i+1}^{n} |\mathrm{Cov}(Z_{n,i}, Z_{n,k})| \\
&\le m + 24 \sum_{i=j+1}^{(j+m)\wedge n} \sum_{k=i+1}^{n} \alpha_{k-i} \|Z_{n,i}\|_\infty \|Z_{n,k}\|_\infty \\
&\le m + 24 \sum_{i=j+1}^{(j+m)\wedge n} \sum_{k=i+1}^{n} \alpha_{k-i} \\
&\le m + 24m \sum_{\ell=1}^{\infty} \alpha_\ell.
\end{aligned}$$

Since $\sum_{\ell=1}^\infty \alpha_\ell < \infty$, we set $M = 1 + 24 \sum_{\ell=1}^\infty \alpha_\ell < \infty$, then

$$\sigma_{n,m}^2 := \sup_{0 \le j \le n-1} \mathrm{E}\Big[\Big(\sum_{i=j+1}^{(j+m)\wedge n} Z_{n,i}\Big)^2\Big] \le Mm.$$

Since $k/n \to 0$ and $p_x > 0$, for all sufficiently large n, we have $k - np_x \le -\frac{np_x}{2}$. Then

$$\begin{aligned}
\mathrm{P}\big(R_{n,k}(x) > r_x + \varepsilon\big) &= \mathrm{P}\Big(\sum_{i=1}^n Z_{n,i} < k - np_x\Big) \\
&\le \mathrm{P}\Big(\Big|\sum_{i=1}^n Z_{n,i}\Big| > np_x - k\Big) \\
&\le \mathrm{P}\Big(\Big|\sum_{i=1}^n Z_{n,i}\Big| > \frac{np_x}{2}\Big).
\end{aligned}$$

We now apply Lemma 1 with $m_n = \lfloor \frac{3}{64} \frac{np_x}{\log n} \rfloor$. Then $m_n \to \infty$ and $m_n/n \to 0$. Therefore for all sufficiently large n, we have $64M \le \frac{4}{3} p_x m_n$ and $\frac{np_x}{2} > 4m_n$, and then the conditions of Lemma 1 are satisfied with $S(n) = 1$ and $\epsilon_n = np_x/2$. Using $\sigma_{n,m_n}^2 \le Mm_n$, we obtain

$$\begin{aligned}
\mathrm{P}\Big(\Big|\sum_{i=1}^n Z_{n,i}\Big| > \frac{np_x}{2}\Big) &\le 4 \exp\left[-\frac{(np_x/2)^2}{64(n/m_n)\sigma_{n,m_n}^2 + \frac{8}{3}(np_x/2)m_n}\right] + 4\frac{n}{m_n}\alpha_{m_n} \\
&\le 4 \exp\left[-\frac{n^2 p_x^2/4}{64Mn + \frac{4}{3}np_x m_n}\right] + 4\frac{n}{m_n}\alpha_{m_n} \\
&\le 4 \exp\left[-\frac{n^2 p_x^2/4}{\frac{8}{3}np_x m_n}\right] + 4\frac{n}{m_n}\alpha_{m_n} \\
&= 4 \exp\left(-\frac{3}{32}\frac{np_x}{m_n}\right) + 4\frac{n}{m_n}\alpha_{m_n}.
\end{aligned}$$

By the choice of m_n, the exponential term is $O(n^{-2})$.

By the assumed bound $\alpha_m \lesssim m^{-\gamma}$ for some $\gamma > 1$,

$$\frac{n}{m_n}\alpha_{m_n} \lesssim \frac{n}{m_n^{1+\gamma}} \lesssim \frac{(\log n)^{1+\gamma}}{n^{\gamma}},$$

which is summable in n. Then

$$\sum_{n=1}^{\infty} \mathrm{P}\big(R_{n,k}(x) > r_x + \varepsilon\big) < \infty.$$

By the Borel–Cantelli lemma,

$$\mathrm{P}\big(R_{n,k}(x) > r_x + \varepsilon \text{ i.o.}\big) = 0.$$

That is, for every fixed $\varepsilon > 0$,

$$R_{n,k}(x) \le r_x + \varepsilon \qquad \text{eventually a.s.}$$

Let $\varepsilon = 1/j$, $j \in \mathbb{N}$, and intersect the resulting probability-one events. Then we obtain

$$\limsup_{n\to\infty} R_{n,k}(x) \le r_x \qquad \text{a.s.}$$

Together with the lower bound $R_{n,k}(x) \ge r_x$ a.s. for every n, we obtain

$$R_{n,k}(x) \to r_x \qquad \text{almost surely.}$$

This completes the proof.

□

A.2 Proof of Theorem 2

Proof. Fix $A > 0$, set $b_n = b_{n,A} := \lceil ((A+1)/\gamma) \log n \rceil$, and suppose $K_A \log n \le k \le \kappa n$, with K_A chosen below. Fix $r > 0$, and define

$$Y_i(r) := \mathbf{1}\{\|X_{n,i} - x\| \le r\}, \qquad p_r := \mathrm{E}[Y_i(r)] = \mu(B(x,r)).$$

Let

$$N_n(x,r) := \sum_{i=1}^{n} Y_i(r).$$

By the definition of the k-nearest-neighbor radius,

$$R_{n,k}(x) > r \iff N_n(x,r) < k.$$

We now use a blocking argument. Partition the row into alternating kept and discarded blocks of length b_n. Set $M_n = \lfloor \frac{n}{2b_n} \rfloor$, and for $u = 1, \dots, M_n$, define the kept block

$$I_u := \{(2u-2)b_n + 1, \dots, (2u-1)b_n\}.$$

Let

$$W_u(r) := \sum_{i \in I_u} Y_i(r), \qquad S_n^{\mathrm{keep}}(r) := \sum_{u=1}^{M_n} W_u(r).$$

Since $S_n^{\mathrm{keep}}(r) \le N_n(x,r)$, we obtain

$$\mathrm{P}(R_{n,k}(x) > r) = \mathrm{P}(N_n(x,r) < k) \le \mathrm{P}(S_n^{\mathrm{keep}}(r) < k).$$

Consider the radius

$$r_j = 2^j r_*(n,k) = 2^j \Big(\frac{8k}{c_- n}\Big)^{1/s}, \qquad j \ge 0,$$

and assume $r_j \le r_0$. By Assumption 2,

$$p_{r_j} = \mu(B(x, r_j)) \ge c_- r_j^s = 8 \cdot 2^{js} \frac{k}{n}.$$

Since $b_n/n \to 0$ and $M_n b_n \geq n/4$ for all sufficiently large n,

$$\mu_j = \mathrm{E}[S_n^{\mathrm{keep}}(r_j)] = M_n b_n p_{r_j} \geq n\, p_{r_j}/4 \geq 2^{js+1}k.$$

Therefore, on the event $\{S_n^{\mathrm{keep}}(r_j) < k\}$, since $\mu_j \geq 2k$,

$$\mu_j - S_n^{\mathrm{keep}}(r_j) > \mu_j - k \geq \mu_j/2.$$

Then

$$\{S_n^{\mathrm{keep}}(r_j) < k\} \subseteq \left\{ \left| S_n^{\mathrm{keep}}(r_j) - \mu_j \right| > \frac{\mu_j}{2} \right\}.$$

Therefore, if we define

$$V_u := W_u(r_j) - \mathrm{E}[W_u(r_j)], \qquad u = 1, \ldots, M_n,$$

we have

$$\mathrm{P}(R_{n,k}(x) > r_j) \leq \mathrm{P}\Big(\Big| \sum_{u=1}^{M_n} V_u \Big| > \frac{\mu_j}{2} \Big).$$

Next we apply Lemma 1 for the block sequence $[V_u]_{u=1}^{M_n}$. Each V_u is measurable with respect to the coordinates in the kept block I_u. For $1 \leq \ell < M_n$ and $1 \leq u \leq M_n - \ell$, the finite-index inclusions

$$\sigma(V_1, \ldots, V_u) \subseteq {}^n\mathcal{F}_1^{(2u-1)b_n},$$
$$\sigma(V_{u+\ell}, \ldots, V_{M_n}) \subseteq {}^n\mathcal{F}_{(2(u+\ell)-2)b_n+1}^n$$

show that the corresponding past and future are separated by $(2\ell-1)b_n + 1$ indices. Thus the mixing coefficients of the kept-block sequence satisfy

$$\bar{\alpha}_\ell \leq \alpha_{(2\ell-1)b_n+1} \leq \alpha_{\ell b_n}, \qquad 1 \leq \ell < M_n,$$

using the monotonicity of the original mixing coefficients. In particular, $\bar{\alpha}_1 \leq \alpha_{b_n} \leq e^{-\gamma b_n} \leq n^{-(A+1)}$ by the choice of b_n. Also, $0 \leq W_u(r_j) \leq b_n$, so $|V_u| = |W_u(r_j) - \mathrm{E}W_u(r_j)| \leq b_n$. Thus the boundedness parameter in Lemma 1 is $S(M_n) = b_n$.

To bound the variance term $\sigma_{M_n,1}^2$, since $W_u(r_j)$ is a sum of b_n Bernoulli indicators, we have $W_u(r_j)^2 \leq b_n W_u(r_j)$, and therefore

$$\mathrm{Var}(V_u) \leq \mathrm{E}[W_u(r_j)^2] \leq b_n\, \mathrm{E}[W_u(r_j)] = b_n^2 p_{r_j}.$$

Then for the block sequence with block parameter $m = 1$,

$$\sigma_{M_n,1}^2 = \sup_{0 \leq v \leq M_n - 1} \mathrm{E}[V_{v+1}^2] \leq b_n^2 p_{r_j}.$$

We now apply Lemma 1 to $[V_u]_{u=1}^{M_n}$ with $m = 1$ and threshold $\varepsilon = \mu_j/2 = M_n b_n p_{r_j}/2$. Since $\varepsilon \geq k \geq K_A \log n$, and $b_n \asymp \log n$, we may choose K_A sufficiently large so that $\varepsilon > 4b_n$ for all sufficiently large n. Then by Lemma 1,

$$\begin{aligned}
\mathrm{P}\Big(\Big| \sum_{u=1}^{M_n} V_u \Big| > \frac{\mu_j}{2} \Big) &\leq 4 \exp\left(- \frac{(\mu_j/2)^2}{64 M_n \sigma_{M_n,1}^2 + \frac{8}{3}(\mu_j/2) b_n} \right) + 4 M_n \bar{\alpha}_1 \\
&\leq 4 \exp\left(- \frac{(\mu_j/2)^2}{64 M_n b_n^2 p_{r_j} + \frac{4}{3}\mu_j b_n} \right) + 4 M_n \bar{\alpha}_1.
\end{aligned}$$

Using $\mu_j = M_n b_n p_{r_j}$, we obtain $\frac{4}{3}\mu_j b_n = \frac{4}{3} M_n b_n^2 p_{r_j}$, so the denominator is bounded by

$$\left(64 + \frac{4}{3}\right) M_n b_n^2 p_{r_j} \leq 66\, M_n b_n^2 p_{r_j},$$

and then

$$\frac{(\mu_j/2)^2}{64 M_n b_n^2 p_{r_j} + \frac{4}{3}\mu_j b_n} \geq c \frac{\mu_j^2}{M_n b_n^2 p_{r_j}} = c \frac{(M_n b_n p_{r_j})^2}{M_n b_n^2 p_{r_j}} = c M_n p_{r_j}.$$

Since $M_n \geq n/(4b_n)$ for all large n,

$$M_n p_{r_j} \geq \frac{n}{4b_n} \cdot 8 \cdot 2^{js} \frac{k}{n} = 2 \cdot 2^{js} \frac{k}{b_n}.$$

Therefore, for some universal constant $c_0 > 0$,

$$\mathrm{P}(R_{n,k}(x) > r_j) \leq 4 \exp\left(-c_0 2^{js} \frac{k}{b_n}\right) + 4M_n \bar{\alpha}_1.$$

Finally, using $b_n \asymp \log n$, $\bar{\alpha}_1 \leq n^{-(A+1)}$, and $M_n \leq n/(2b_n)$, we obtain $4M_n\bar{\alpha}_1 \leq C_A n^{-A}$, and then

$$\mathrm{P}(R_{n,k}(x) > r_j) \leq 4 \exp\left(-c_A\, 2^{js} \frac{k}{\log n}\right) + C_A n^{-A},$$

where c_A, C_A, K_A depend only on A and γ. Compact support was not used. This completes the proof. □

A.3 Proof of Theorem 3

Proof. Fix $A > p/s$ and choose $K_0 \geq K_A$ as in Theorem 2; thus K_0 may depend on p. Let

$$r_* = \Big(\frac{8k}{c_- n}\Big)^{1/s}$$

and choose J_0 maximal such that $2^{J_0} r_* \leq r_0$. Assumption 2 implies $x \in \mathrm{supp}(\mu)$. Since the support of μ has diameter at most D, we have $R_{n,k}(x) \leq D$ almost surely.

We decompose the moment over

$$\{R_{n,k}(x) \leq r_*\}, \quad \{2^j r_* < R_{n,k}(x) \leq 2^{j+1} r_*\},\ 0 \leq j \leq J_0 - 1, \quad \{R_{n,k}(x) > 2^{J_0} r_*\},$$

and we obtain

$$\mathrm{E}[R_{n,k}(x)^p] \leq r_*^p + \sum_{j=0}^{J_0-1} (2^{j+1} r_*)^p\, \mathrm{P}(R_{n,k}(x) > 2^j r_*) + D^p\, \mathrm{P}(R_{n,k}(x) > 2^{J_0} r_*).$$

We now apply Theorem 2. For every $0 \leq j \leq J_0$,

$$\mathrm{P}(R_{n,k}(x) > 2^j r_*) \leq 4 \exp\Big(-c_A\, 2^{js} \frac{k}{\log n}\Big) + Cn^{-A}.$$

At $j = J_0$, maximality of J_0 gives

$$2^{J_0 s} k = \frac{c_- n}{8} (2^{J_0} r_*)^s > \frac{c_- n}{8} (r_0/2)^s.$$

Consequently, the exponential term at J_0 is at most $4e^{-cn/\log n} \leq Cn^{-A}$ for all sufficiently large n. We therefore obtain

$$\mathrm{E}[R_{n,k}(x)^p] \leq r_*^p + C r_*^p \sum_{j=0}^{J_0-1} 2^{jp} \exp\Big(-c_A\, 2^{js} \frac{k}{\log n}\Big) + C r_*^p n^{-A} \sum_{j=0}^{J_0-1} 2^{jp} + C D^p n^{-A}.$$

We now bound the three terms on the right-hand side. For the first term, since $k \geq K_0 \log n$, the quantity $k/\log n$ is bounded below by a positive constant. Therefore

$$\sum_{j=0}^{J_0-1} 2^{jp} \exp\Big(-c_A\, 2^{js} \frac{k}{\log n}\Big) \leq \sum_{j \geq 0} 2^{jp} e^{-c_1 2^{js}} < \infty$$

for some constant $c_1 > 0$. For the second term, since $2^{J_0} r_* \leq r_0 < 2^{J_0+1} r_*$, we have $2^{J_0} \asymp r_*^{-1}$, where the constants depend only on r_0. Then

$$r_*^p \sum_{j=0}^{J_0-1} 2^{jp} \leq C r_*^p 2^{J_0 p} \leq C,$$

and then

$$Cr_*^p n^{-A} \sum_{j=0}^{J_0-1} 2^{jp} \le Cn^{-A}.$$

Since $A > p/s$ and $k \ge 1$, we have $n^{-A} \le n^{-p/s} \le (k/n)^{p/s} \asymp r_*^p$. The third term $CD^p n^{-A}$ is bounded in the same way. Combining the above estimates yields

$$\mathrm{E}[R_{n,k}(x)^p] \le Cr_*^p + Cn^{-A} \le Cr_*^p.$$

Since

$$r_*^p = \Big(\frac{8k}{c_- n}\Big)^{p/s} \asymp \Big(\frac{k}{n}\Big)^{p/s},$$

we obtain

$$\mathrm{E}[R_{n,k}(x)^p] \le C\Big(\frac{k}{n}\Big)^{p/s}.$$

This completes the proof. □

A.4 Proof of Theorem 4

Proof. Let

$$r^- = \Big(\frac{k}{2c_+ n}\Big)^{1/s}.$$

By Assumption 3,

$$\mu(B(x, r^-)) \le c_+ (r^-)^s = \frac{k}{2n}.$$

Therefore,

$$\mathrm{E}[N_n(x, r^-)] = n\,\mu(B(x, r^-)) \le k/2,$$

where,

$$N_n(x, r) = \sum_{i=1}^{n} \mathbf{1}\{\|X_{n,i} - x\| \le r\}.$$

By Markov's inequality,

$$\mathrm{P}\big(N_n(x, r^-) \ge k\big) \le \frac{\mathrm{E}[N_n(x, r^-)]}{k} \le \frac{1}{2},$$

and

$$R_{n,k}(x) > r^- \iff N_n(x, r^-) < k,$$

we have

$$\mathrm{P}\big(R_{n,k}(x) > r^-\big) \ge \frac{1}{2}.$$

On the event $\{R_{n,k}(x) > r^-\}$, we have $R_{n,k}(x)^p \ge (r^-)^p$, and then

$$\mathrm{E}[R_{n,k}(x)^p] \ge (r^-)^p\,\mathrm{P}\big(R_{n,k}(x) > r^-\big) \ge \frac{1}{2}\,(r^-)^p.$$

By the definition of r^-, we obtain

$$\mathrm{E}[R_{n,k}(x)^p] \ge \frac{1}{2}\Big(\frac{k}{2c_+ n}\Big)^{p/s}.$$

This completes the proof. □

A.5 Proof of Theorem 5

Proof. The conclusion follows directly by combining the upper and lower bounds established in Theorems 3 and 4. □

Table 3: Dependence-strength settings for the three dependent DGPs in experiment 1. Weak settings are near i.i.d., while medium and strong settings increase dependence contrast.

Model	Tuning parameter	Weak	Medium	Strong
Linear Gaussian state-space (factor AR(1))	AR coefficient ρ	0.02	0.60	0.98
Hidden Markov model (2-state)	Stay probability p_{stay}	0.70	0.99	0.9995
Gaussian process (FFT, RBF)	Lengthscale ℓ	1	50	200

B Experiments

We provide additional details on the implementation of the experiments.

B.1 Experiment 1 (Section 4.1) implementation details

Setup. Experiment 4.1 uses interior query aggregation. For each MC replication, we sample with the global random seed 20260504

$$X_{\text{eval}} \sim \text{Unif}([0.01, 0.99]^d), \qquad |X_{\text{eval}}| = 1000,$$

The number of NNs is chosen as $k = \lfloor n^\beta \rceil$ with β on a 20-point grid in $[0.1, 0.9]$ for slope fitting. We also impose a cap for $\frac{k}{n} \leq 0.01$.

Data generating processes. The i.i.d. uniform data is used as the baseline in every panel. The dependent data is generated by the following processes, with three strength levels, as shown in Table 3.

- **Linear Gaussian state-space (LSS).** A shared latent AR(1) factor is generated as
$$f_t = \rho f_{t-1} + \sqrt{1-\rho^2}\,\eta_t, \qquad \eta_t \sim \mathcal{N}(0,1).$$
Then each coordinate is
$$z_{t,j} = a f_t + \sqrt{1-a^2}\,\varepsilon_{t,j}, \qquad \varepsilon_{t,j} \sim \mathcal{N}(0,1),$$
with dimension-damped shared loading
$$a = \frac{s\sqrt{|\rho|}}{d^\gamma}, \qquad s = 0.9,\ \gamma = 0.3.$$
(Implementation clips $a \in [0, 0.999]$ and uses burn-in 400.)
- **Hidden Markov model (HMM).** Each coordinate j has its own two-state chain $S_{t,j} \in \{0, 1\}$ with
$$\Pr(S_{t,j} = S_{t-1,j}) = p_{\text{stay}}.$$
Conditional emissions are
$$Z_{t,j} \mid S_{t,j} \sim \mathcal{N}\big((2S_{t,j} - 1)\mu_{\text{eff}},\, \sigma^2\big),$$
where $\mu = 1.75$, $\sigma = 0.70$, and
$$\mu_{\text{eff}} = \mu\,|2p_{\text{stay}} - 1|^6.$$
(This keeps weak settings close to i.i.d. after marginal transformation while separating medium/strong regimes; burn-in 400.)
- **Gaussian process (FFT, RBF).** For each coordinate, a stationary Gaussian sequence is generated with lag covariance
$$k(h) = \exp\left(-\frac{h^2}{2\ell^2}\right), \qquad \ell \in \{1, 50, 200\},$$
using circulant embedding and FFT spectral filtering ($O(n \log n)$), then truncating to length n.
- **Transformation.** All d coordinates are transformed to have approximately $\text{Unif}[0, 1]$ marginal distributions while retaining dependence:
$$U_{t,j} = \Phi(Z_{t,j}) \quad \text{(LSS/GP)},$$
and, for HMM, via the stationary two-component mixture CDF
$$F_{\text{mix}}(z) = \tfrac{1}{2}\Phi\left(\frac{z + \mu_{\text{eff}}}{\sigma}\right) + \tfrac{1}{2}\Phi\left(\frac{z - \mu_{\text{eff}}}{\sigma}\right).$$

B.2 Experiment 2 (Section 4.2) implementation details

Data-generating process. For each intrinsic dimension s, we generate independent stationary Gaussian AR(1) latent processes across coordinates:

$$Z_t = \rho Z_{t-1} + \sqrt{1-\rho^2}\,\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, I_s).$$

The temporal dependence levels are $\rho = 0$ for the i.i.d. benchmark, $\rho = 0.3$ for weak AR(1) dependence, and $\rho = 0.6$ for strong AR(1) dependence. The latent coordinates are independent across dimensions, while each coordinate follows the same AR(1) dependence over time. The latent vector is embedded into a 5-dimensional signal representation and then mapped into 20 observed coordinates by a fixed orthonormal 20×5 linear map, which induces a low-rank cross-coordinate dependence structure in the ambient observations. The PCA estimator fits a 5-component PCA representation to the observed data and then estimates entropy using the top s PCA coordinates, which, in this linear Gaussian setting, recover an orthogonal low-dimensional representation up to rotation.

Estimators and simulation grid. All reported curves use the Kozachenko–Leonenko k-NN entropy estimator with Euclidean distance. The oracle intrinsic estimate is computed on the true latent s-dimensional sample, and the PCA estimate is computed on the top s PCA coordinates. The current implementation uses 1000 Monte Carlo replications with global random seed 20260504, burn-in 2000, and neighbor schedule $k = \max\{2, \lceil n^{0.1} \rceil\}$.

B.3 Verification of Assumptions 1–4 for the synthetic experiments: proofs

This appendix verifies that the data-generating processes used in the experiments satisfy the assumptions of the theory: we prove that each Experiment-1 generator is geometrically α-mixing with an explicit rate and explicit constants (Assumption 1; for the GP the bound is in fact super-geometric), and that its marginal law satisfies the local mass conditions with exponent $s = d$ and explicit constants (Assumptions 2–3); Assumption 4 is immediate because all samples take values in $[0,1]^d$. We also show (Proposition B.6) that the compact-support assumption in Theorem 3 can be replaced by a marginal moment condition, which covers the Gaussian design of Experiment 2. Appendix B.4 complements these proofs with a empirical diagnostics on the simulation generators; these complement rather than replace the proofs.

Throughout, Φ and ϕ denote the standard normal distribution and density functions, $\omega_d := \pi^{d/2}/\Gamma(d/2+1)$ the volume of the d-dimensional unit ball, and $\lambda_2 := 2p_{\mathrm{stay}} - 1$ the second eigenvalue of the two-state symmetric transition matrix. We use the convention $\log(1/0) = \infty$; when $\rho = 0$ or $\lambda_2 = 0$, the corresponding process is independent across time and Assumption 1 holds for every finite $\gamma > 0$. All three Experiment-1 generators are *stationary* sequences initialized from their stationary laws (the implementation adds a burn-in, which is then redundant); the triangular array of Section 2 is obtained by the sequential embedding $X_{n,i} = X_i$, so the array coefficients α_m coincide with the sequence coefficients.

Recall the three generators. Fix $d \geq 1$. In each case the observed vector is $X_t \in [0,1]^d$.

- **Linear state space (LSS).** A stationary Gaussian AR(1) factor $f_t = \rho f_{t-1} + \sqrt{1-\rho^2}\,\eta_t$ with $f_0 \sim N(0,1)$, $\eta_t \overset{\text{iid}}{\sim} N(0,1)$, and observations $X_{t,j} = \Phi\big(a f_t + \sqrt{1-a^2}\,\varepsilon_{t,j}\big)$, $j = 1, \ldots, d$, with $\varepsilon_{t,j} \overset{\text{iid}}{\sim} N(0,1)$ independent of $[f_t]$ and loading $a = 0.9\sqrt{|\rho|}/d^{0.3} \in [0,1)$.
- **Hidden Markov model (HMM).** Per coordinate j an independent stationary two-state chain $S_{t,j} \in \{0,1\}$ with stay probability p_{stay} and uniform initial law, emissions $z_{t,j} = \mu_{\mathrm{eff}}(2S_{t,j} - 1) + \sigma\xi_{t,j}$ with $\xi_{t,j} \overset{\text{iid}}{\sim} N(0,1)$, and $X_{t,j} = F(z_{t,j})$, where $F(z) = \frac{1}{2}\Phi((z + \mu_{\mathrm{eff}})/\sigma) + \frac{1}{2}\Phi((z - \mu_{\mathrm{eff}})/\sigma)$ is the exact stationary emission distribution function.
- **Gaussian process (GP–RBF).** Per coordinate j an independent stationary Gaussian sequence $[w_{t,j}]_t$ with autocovariance $r_h = e^{-h^2/(2\ell^2)}$ and $X_{t,j} = \Phi(w_{t,j})$.*

*The implementation samples the finite window by circulant embedding, which reproduces the covariance r_h within the window exactly whenever the embedding is nonnegative definite; negative eigenvalues, when they occur, are at the level of numerical noise and are clipped at zero.

Preliminaries We collect standard facts about the dependence coefficients $\alpha(\mathcal{A},\mathcal{B})$, $\beta(\mathcal{A},\mathcal{B})$, and the maximal correlation $\rho(\mathcal{A},\mathcal{B})$ between two σ-fields.

Lemma 2. *Let $[U_t]$, $[V_t]$ be random sequences on a common probability space.*

(i) (Instantaneous functions.) If $Y_t = g(U_t)$ for a fixed measurable g, then $\alpha_Y(m) \le \alpha_U(m)$ for all m, and similarly for β.

(ii) (Independent components.) If $[U_t]$ and $[V_t]$ are independent of each other, then the paired sequence $W_t = (U_t, V_t)$ satisfies $\alpha_W(m) \le \alpha_U(m)+\alpha_V(m)$, $\beta_W(m) \le \beta_U(m)+\beta_V(m)$, and $\rho_W(m) = \max\{\rho_U(m), \rho_V(m)\}$ [9, Theorem 5.1(a)–(c)].

(iii) (Comparison.) For any two σ-fields, $2\alpha(\mathcal{A},\mathcal{B}) \le \beta(\mathcal{A},\mathcal{B})$ and $4\alpha(\mathcal{A},\mathcal{B}) \le \rho(\mathcal{A},\mathcal{B})$ [9, eqs. (1.11)–(1.12)].

(iv) (Gaussian sequences.) If $[U_t]$ is a stationary mean-zero Gaussian sequence, its maximal correlation coefficient equals its maximal linear *correlation coefficient: $\rho_U(m) = \sup|\mathrm{Corr}(L_1, L_2)|$, the supremum running over L_1 in the closed linear span of $[U_s : s \le t]$ and L_2 in that of $[U_s : s \ge t+m]$ [9, eq. (7.1)], going back to [33].*

Part (i) is immediate from $\sigma(Y_i : i \in I) \subseteq \sigma(U_i : i \in I)$. We also record how a multiplicative constant is absorbed into the geometric rate, so that bounds of the form $\alpha_m \le Ce^{-\gamma m}$ with $C \ge 1$ imply Assumption 1 verbatim.

Remark 2 (Absorbing constants)**.** *Suppose $\alpha_m \le Ce^{-\gamma m}$ for all $m \ge 1$ with $C \ge 1$, $\gamma > 0$. Since always $\alpha_m \le 1/4$, setting $m_0 := \lceil \gamma^{-1}\log(4C)\rceil$ gives $\alpha_m \le \frac{1}{4}e^{-\gamma(m-m_0)_+}$, and hence $\alpha_m \le e^{-\gamma' m}$ for all $m \ge 1$ with $\gamma' := \min\{\gamma,\ \log(4)/m_0\} > 0$. Thus geometric decay with any constant yields Assumption 1 after a benign adjustment of γ.*

Assumption 1: geometric α-mixing

Proposition B.1 (LSS)**.** *For any $d \ge 1$, $\rho \in (-1,1)$, and $a \in [0,1)$, the LSS sequence satisfies*

$$\alpha_m \ \le\ \tfrac{1}{4}\,|\rho|^m \qquad \text{for all } m \ge 1,$$

so Assumption 1 holds with $\gamma = \log(1/|\rho|)$.

Proof. Write $X_t = G(f_t, \varepsilon_t)$ with $G(f,e) = (\Phi(af + \sqrt{1-a^2}\,e_j))_{j\le d}$ and $\varepsilon_t = (\varepsilon_{t,1},\dots,\varepsilon_{t,d})$. By Lemma 2(i), $\alpha_X(m) \le \alpha_{(f,\varepsilon)}(m)$. Since $[\varepsilon_t]$ is i.i.d. and independent of $[f_t]$, Lemma 2(ii) gives $\alpha_{(f,\varepsilon)}(m) \le \alpha_f(m) + 0$.

It remains to bound $\alpha_f(m)$ for the stationary Gaussian AR(1) factor, for which $\mathrm{Cov}(f_t, f_{t+h}) = \rho^{|h|}$. By Lemma 2(iv) it suffices to bound correlations of linear combinations. Let L_1 lie in the closed linear span of $[f_s : s \le t]$ and $L_2 = \sum_j v_j f_{t+m+j}$ (finite sum, $j \ge 0$). By the Markov property and Gaussianity, $\mathrm{E}[f_{t+m+j} \mid f_s,\ s \le t] = \rho^{m+j} f_t$, so with $b := \sum_j v_j \rho^j$,

$$\mathrm{Cov}(L_1, L_2) = \mathrm{Cov}\big(L_1,\ \rho^m b\, f_t\big) \le |\rho|^m\,|b|\,\sqrt{\mathrm{Var}(L_1)}.$$

Moreover $\mathrm{Var}(L_2) \ge \mathrm{Var}\big(\mathrm{E}[L_2 \mid f_{t+m}]\big) = b^2$, whence $|\mathrm{Corr}(L_1, L_2)| \le |\rho|^m$. Taking suprema (and closing in L^2) yields $\rho_f(m) = |\rho|^m$, and Lemma 2(iii) gives $\alpha_f(m) \le |\rho|^m/4$. □

For the three strength levels $\rho \in \{0.02, 0.60, 0.98\}$ this gives $\gamma \approx 3.91, 0.51, 0.020$, i.e. dependence horizons $1/\gamma \approx 0.3, 2, 50$ steps.

Proposition B.2 (HMM)**.** *For any $d \ge 1$ and $p_{\mathrm{stay}} \in (0,1)$, the HMM sequence satisfies*

$$\alpha_m \ \le\ \tfrac{1}{4}\,|\lambda_2|^m, \qquad \lambda_2 = 2p_{\mathrm{stay}} - 1,$$

so Assumption 1 holds with $\gamma = \log(1/|\lambda_2|)$.

Proof. As before, $X_t = G(S_t, \xi_t)$ with ξ_t i.i.d. standard normal and independent of the latent chains, so by Lemma 2(i)–(ii), $\alpha_X(m) \le \alpha_S(m)$, where $S_t = (S_{t,1},\dots,S_{t,d})$ collects the d independent stationary two-state chains.

We first bound the maximal correlation of a single chain with transition matrix $T = \left(\begin{smallmatrix} p & 1-p \\ 1-p & p \end{smallmatrix}\right)$, $p = p_{\mathrm{stay}}$. The chain is reversible with respect to $\pi = (\frac{1}{2}, \frac{1}{2})$, so T is self-adjoint on $L^2(\pi)$ with eigenvalues 1 and $\lambda_2 = 2p - 1$; hence for every h with $\pi h = 0$,

$$\|T^m h\|_{L^2(\pi)} \;\le\; |\lambda_2|^m \, \|h\|_{L^2(\pi)}.$$

Let $f \in L^2(\sigma(S_u : u \le t))$ and $g \in L^2(\sigma(S_u : u \ge t+m))$ be centered. By the Markov property, $\mathrm{E}[g \mid S_{t+m}] =: h(S_{t+m})$ with $\pi h = 0$ and $\mathrm{Var}(h(S_{t+m})) \le \mathrm{Var}(g)$, and $\mathrm{E}[g \mid S_u,\ u \le t] = (T^m h)(S_t)$. Therefore, by Cauchy–Schwarz and stationarity,

$$\mathrm{Cov}(f, g) = \mathrm{E}\big[f\,(T^m h)(S_t)\big] \le \sqrt{\mathrm{Var}(f)}\,\|T^m h\|_{L^2(\pi)} \le |\lambda_2|^m \sqrt{\mathrm{Var}(f)\,\mathrm{Var}(g)},$$

i.e. $\rho_{\mathrm{chain}}(m) \le |\lambda_2|^m$. By Lemma 2(ii), the d independent chains satisfy $\rho_S(m) = \rho_{\mathrm{chain}}(m) \le |\lambda_2|^m$, and Lemma 2(iii) gives $\alpha_S(m) \le \frac{1}{4}\rho_S(m) \le \frac{1}{4}|\lambda_2|^m$.

(Alternatively, Davydov's identity $\beta(m) = \mathrm{E}_\pi \|T^m(S_0, \cdot) - \pi\|_{\mathrm{TV}}$ for stationary Markov chains [15], together with $T^m(s, \cdot) = \pi + \lambda_2^m(\delta_s - \pi)$ and the subadditivity in Lemma 2(ii), gives the slightly weaker bound $\alpha_m \le \frac{d}{4}|\lambda_2|^m$.) □

For $p_{\mathrm{stay}} \in \{0.70, 0.99, 0.9995\}$: $\lambda_2 \in \{0.4, 0.98, 0.999\}$ and $\log(1/\lambda_2) \approx 0.92, 0.020, 0.0010$, i.e. horizons $\approx 1, 50, 1000$ steps. Note that the strong setting is *still geometrically mixing*; its horizon is simply comparable to the sample sizes used in Experiment 1, which is the intended stress regime.

Proposition B.3 (GP–RBF). *Fix $\ell > 0$ and $d \ge 1$, and set*

$$f_{*,\ell} := \sqrt{2\pi}\,\ell\, e^{-\pi^2\ell^2/2}, \qquad K_\ell := \frac{1}{f_{*,\ell}\big(1 - e^{-2/\ell^2}\big)}.$$

The GP–RBF sequence satisfies

$$\alpha_m \;\le\; \tfrac{1}{4}\, K_\ell\, e^{-m^2/(2\ell^2)} \qquad \textit{for all } m \ge 1.$$

In particular the decay is super-geometric, and Assumption 1 holds for every fixed ℓ with $\gamma(\ell) = \min\big\{m_\ell/\ell^2,\ \log(4)/m_\ell\big\} > 0$, where $m_\ell := \big\lceil \ell\sqrt{2\log(4K_\ell)}\,\big\rceil$.

Proof. Step 1 (spectral floor). Fix one coordinate and write w_t for it. Its spectral density is, by Poisson summation,

$$f_\ell(\lambda) \;=\; \sum_{h\in\mathbb{Z}} e^{-h^2/(2\ell^2)} e^{-ih\lambda} \;=\; \sqrt{2\pi}\,\ell \sum_{k\in\mathbb{Z}} \exp\Big(-\tfrac{\ell^2(\lambda+2\pi k)^2}{2}\Big), \qquad \lambda \in [-\pi, \pi],$$

and all terms are positive, so keeping only $k = 0$ gives $f_\ell(\lambda) \ge f_{*,\ell}$ on $[-\pi, \pi]$. Hence, for any finitely supported real $(a_u)_{u\ge0}$ and $U := \sum_{u\ge0} a_u w_{t-u}$, writing $A(\lambda) := \sum_{u\ge0} a_u e^{iu\lambda}$, Parseval's identity gives

$$\mathrm{Var}(U) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |A(\lambda)|^2 f_\ell(\lambda)\, d\lambda \;\ge\; f_{*,\ell}\,\|a\|_2^2,$$

and likewise $\mathrm{Var}(V) \ge f_{*,\ell}\,\|b\|_2^2$ for $V := \sum_{v\ge0} b_v w_{t+m+v}$.

Step 2 (Hankel bound on the cross-covariance). With $r_h = e^{-h^2/(2\ell^2)}$,

$$\mathrm{Cov}(U, V) = \sum_{u,v\ge0} a_u b_v\, r_{m+u+v} = \langle a,\ H_m b\rangle, \qquad (H_m)_{uv} := r_{m+u+v},$$

so $|\mathrm{Cov}(U,V)| \le \|H_m\|_{\mathrm{op}}\|a\|_2\|b\|_2 \le \|H_m\|_{\mathrm{HS}}\|a\|_2\|b\|_2$. Counting the $q+1$ pairs with $u+v = q$ and using $(m+q)^2 \ge m^2 + 2mq$ and $m \ge 1$,

$$\begin{aligned}\|H_m\|_{\mathrm{HS}}^2 &= \sum_{q\ge0}(q+1)\, e^{-(m+q)^2/\ell^2} \le e^{-m^2/\ell^2}\sum_{q\ge0}(q+1)\, e^{-2mq/\ell^2} \\ &= \frac{e^{-m^2/\ell^2}}{\big(1 - e^{-2m/\ell^2}\big)^2} \le \frac{e^{-m^2/\ell^2}}{\big(1 - e^{-2/\ell^2}\big)^2}.\end{aligned}$$

Combining the two steps, $|\mathrm{Corr}(U,V)| \le K_\ell\, e^{-m^2/(2\ell^2)}$; the bound is uniform over finite combinations and hence extends to the closed linear spans. By Lemma 2(iv), $\rho_w(m) \le K_\ell\, e^{-m^2/(2\ell^2)}$.

Step 3 (coordinates and transform). The d coordinate sequences are i.i.d. copies, so Lemma 2(ii) gives $\rho_W(m) = \rho_w(m)$ for the vector sequence $W_t = (w_{t,1}, \dots, w_{t,d})$; the coordinatewise transform Φ is instantaneous, so by Lemma 2(i) and (iii), $\alpha_X(m) \le \alpha_W(m) \le \frac{1}{4}\rho_W(m) \le \frac{1}{4}K_\ell e^{-m^2/(2\ell^2)}$.

Step 4 (explicit geometric exponent). By definition of m_ℓ, $K_\ell e^{-m_\ell^2/(2\ell^2)} \le \frac{1}{4}$. For $m \ge m_\ell$, using $m^2 - m_\ell^2 \ge 2m_\ell(m - m_\ell)$,

$$\alpha_m \le \tfrac{1}{4}K_\ell e^{-m^2/(2\ell^2)} \le \tfrac{1}{16}\, e^{-m_\ell(m-m_\ell)/\ell^2},$$

while $\alpha_m \le \frac{1}{4}$ always. Exactly as in Remark 2, these two bounds give $\alpha_m \le e^{-\gamma(\ell)m}$ for all $m \ge 1$ with $\gamma(\ell) = \min\{m_\ell/\ell^2, \log(4)/m_\ell\}$. □

Remark 3 (Sharpness of the GP constants)**.** *For $\ell \ge 1$, $\log K_\ell = \pi^2\ell^2/2 + O(\log \ell)$, so $m_\ell \approx \pi\ell^2$ and $\gamma(\ell) \asymp \ell^{-2}$: the provable bound in Proposition B.3 activates only beyond $m \approx \pi\ell^2$. This conservativeness comes from the worst-case spectral floor $f_{*,\ell}$; at the level of pairs of observations the dependence is far smaller,* $\mathrm{Corr}(w_t, w_{t+m}) = e^{-m^2/(2\ell^2)}$*, i.e. negligible beyond $m \approx \ell$, which matches the measured integrated autocorrelation times in Table 4. The GP model is included in Experiment 1 precisely as a stress test beyond the cleanly parameterized geometric-mixing models.*

Assumptions 2–3: local mass, and Assumption 4

Proposition B.4 (Exactly uniform marginals)**.** *For the HMM and GP–RBF generators with any $d \ge 1$ and any parameter value, and for the LSS generator with $d = 1$, the common marginal law is exactly $\mu = \mathrm{Unif}([0,1]^d)$. Consequently, for every $x \in (0,1)^d$, Assumptions 2 and 3 hold with*

$$s = d, \qquad c_- = c_+ = \omega_d, \qquad r_0 = \mathrm{dist}(x, \partial[0,1]^d),$$

and Assumption 4 holds with $D = \sqrt{d}$.

Proof. Each scalar sequence is stationary with a continuous, strictly increasing marginal distribution function, and the implementation applies exactly that function. For the LSS, $af_t + \sqrt{1-a^2}\,\varepsilon_{t,j} \sim N(0,1)$ because $f_t \sim N(0,1)$ at stationarity and $a^2 + (1-a^2) = 1$; the transform is Φ. For the HMM, $S_{t,j}$ is uniform at stationarity, so $z_{t,j}$ has the two-component mixture distribution with distribution function F above, and the transform is F. For the GP, $w_{t,j} \sim N(0,1)$ since $r_0 = 1$, and the transform is Φ. In each case the probability integral transform yields $X_{t,j} \sim \mathrm{Unif}(0,1)$ exactly.

For the HMM and the GP the d coordinate sequences are independent by construction (independent latent chains and noises; independent columns), so μ is the product of uniform marginals, i.e. Lebesgue measure on $[0,1]^d$; for $d = 1$ this holds trivially for the LSS as well. Finally, for $x \in (0,1)^d$ and $r \le \mathrm{dist}(x, \partial[0,1]^d)$ we have $B(x,r) \subseteq [0,1]^d$ and hence $\mu(B(x,r)) = \mathrm{vol}(B(x,r)) = \omega_d r^d$. □

Proposition B.5 (LSS with $d \ge 2$: Gaussian copula)**.** *For the LSS generator with $d \ge 2$, the marginal law μ is the Gaussian copula measure with equicorrelation matrix $\Sigma = (1-a^2)I_d + a^2\mathbf{1}\mathbf{1}^\top$, whose density*

$$c_\Sigma(u) \;=\; \frac{\phi_\Sigma\big(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\big)}{\prod_{j=1}^d \phi\big(\Phi^{-1}(u_j)\big)}, \qquad u \in (0,1)^d,$$

is continuous and strictly positive on the open cube (for $a^2 < 1$, Σ is nonsingular with eigenvalues $1 - a^2$ and $1 + (d-1)a^2$). Consequently, for every $x \in (0,1)^d$, Assumptions 2 and 3 hold with $s = d$: setting $r_0 = r_0(x) = \frac{1}{2}\,\mathrm{dist}(x, \partial[0,1]^d)$ and $K_x = \overline{B(x, r_0)} \subset (0,1)^d$,

$$c_- = \omega_d \min_{K_x} c_\Sigma > 0, \qquad c_+ = \omega_d \max_{K_x} c_\Sigma < \infty,$$

and for all $r \le r_0$, $c_- r^d \le \mu(B(x,r)) = \int_{B(x,r)} c_\Sigma(u)\, du \le c_+ r^d$.

Proof. The joint law of $(z_{t,1}, \dots, z_{t,d})$ is $N_d(0, \Sigma)$ with $\mathrm{Corr}(z_{t,j}, z_{t,k}) = a^2$ for $j \ne k$, so $X_t = (\Phi(z_{t,1}), \dots, \Phi(z_{t,d}))$ has the stated copula law; the density formula and its continuity and positivity

on $(0,1)^d$ are standard for nonsingular Σ. The compact set K_x lies in the open cube, so c_Σ attains a positive minimum and finite maximum on it, and for $r \le r_0$ the ball $B(x,r)$ is contained in $K_x \subseteq [0,1]^d$; the displayed bounds follow by integrating the density over $B(x,r)$. □

Propositions B.4 and B.5 make the division of labor in the theory concrete: the temporal dependence deforms the *joint* law across time (Assumption 1) but the *local mass exponent* of the marginal remains $s = d$ at every dependence level; for the LSS with $d \ge 2$ only the constants $c_\pm$ become query-dependent. Since the slope diagnostic of Experiment 1 estimates the exponent, not the constants, the predicted slope p/d is unaffected. This is confirmed empirically in Appendix B.4 (Table 4): the fitted local-mass exponents satisfy $\hat{s} \in [1.000, 1.001]$ (target $d = 1$) and $\hat{s} \in [2.976, 3.007]$ (target $d = 3$) across all models and strengths.

Remark 4 (Experiment 2)**.** *The Experiment-2 process is the stationary Gaussian AR(1) latent vector $Z_t = \rho Z_{t-1} + \sqrt{1-\rho^2}\,\varepsilon_t \in \mathbb{R}^s$ with independent coordinates, observed as $Y_t = Q\,(Z_t^\top, 0, \dots, 0)^\top \in \mathbb{R}^{20}$ for a fixed orthonormal $Q \in \mathbb{R}^{20\times 5}$ ($Q^\top Q = I_5$). (i)* Mixing: *each coordinate of Z_t is the AR(1) factor of Proposition B.1, whose maximal correlation is ρ^m; since the coordinates are independent, Lemma 2(ii) gives $\rho_Z(m) = \rho^m$, hence $\alpha_m \le \frac{1}{4}\rho^m$ by Lemma 2(iii), and Assumption 1 holds with $\gamma = \log(1/\rho)$ for $\rho \in \{0.3, 0.6\}$; Y_t is an instantaneous function of Z_t, so it inherits the bound. (ii)* Local mass with intrinsic exponent: *Q is an isometry onto its range, so $\|Y_t - y\| = \|Z_t - z\|$ for query points $y = Q(z^\top, 0, \dots, 0)^\top$ on the support, and hence $\mu_Y(B(y,r)) = \mathrm{P}(\|Z - z\| \le r) = \int_{B_s(z,r)} \phi_s(v)\,dv$, where ϕ_s is the $N(0, I_s)$ density. For $r \le 1$ this is bounded between $\omega_s r^s \min_{B_s(z,1)} \phi_s$ and $\omega_s r^s \max_{B_s(z,1)} \phi_s$, so Assumptions 2–3 hold with exponent s — the* intrinsic *dimension — even though the ambient dimension is* 20*. This is exactly the mechanism that the PCA entropy experiment probes. (iii)* Compact support: *Assumption 4 fails as stated (Gaussian support is unbounded), but the Gaussian marginal has finite moments of every order, so Proposition B.6 below applies with any $q > p$: the moment upper bound of Theorem 3 holds verbatim for the Experiment-2 design, with no additional logarithmic factors. Claims (i) and (ii) are illustrated by empirical diagnostics on the Experiment-2 generator in Appendix B.4 (Figure 9).*

Relaxing the compact-support assumption Assumption 4 enters the proof of Theorem 3 only through the deterministic bound $R_{n,k}(x) \le D$ on the far-tail event. The following proposition replaces it by a marginal moment condition.

Proposition B.6 (Moments without compact support)**.** *Assume Assumptions 1 and 2 at x, and suppose that $M_q := \mathrm{E}\|X_{n,1} - x\|^q < \infty$ for some $q > p$, where $p > 0$ is the moment order of interest. Then there exist constants K_0' and C', depending only on $(p, q, s, \gamma, c_-, r_0, M_q)$, such that*

$$\mathrm{E}\big[R_{n,k}(x)^p\big] \;\le\; C'\Big(\frac{k}{n}\Big)^{p/s} \qquad \textit{whenever } K_0' \log n \le k \le \kappa n.$$

Proof. Fix

$$A \;\ge\; A_0 := \frac{p/s + p/q}{1 - p/q} \qquad (\text{note } A_0 \ge p/s).$$

Step 1 (tail bound with remainder n^{-A}). Apply Theorem 2 with the chosen A and $K_0' \ge K_A$. With $r_* = (8k/(c_- n))^{1/s}$ and J_0 maximal such that $2^{J_0} r_* \le r_0$, it gives

$$\mathrm{P}\big(R_{n,k}(x) > 2^j r_*\big) \le 4\exp\Big(-c_A\, 2^{js}\,\frac{k}{\log n}\Big) + C_A n^{-A}, \qquad 0 \le j \le J_0.$$

Compact support is not required for this tail bound.

Step 2 (local dyadic part). Exactly as in the proof of Theorem 3,

$$\mathrm{E}\big[R_{n,k}(x)^p\,\mathbf{1}\{R_{n,k}(x) \le 2^{J_0} r_*\}\big] \le r_*^p + \sum_{j=0}^{J_0 - 1} (2^{j+1} r_*)^p\,\mathrm{P}\big(R_{n,k}(x) > 2^j r_*\big) \le C r_*^p + C n^{-A},$$

using $\sum_{j\ge 0} 2^{jp} e^{-c_A 2^{js} K_0'} < \infty$ for the exponential part and $r_*^p \sum_{j<J_0} 2^{jp} \le C(r_0, p)$ (from $2^{J_0} \asymp r_0/r_*$) for the remainder part. Since $A \ge p/s$ and $k \ge 1$, $n^{-A} \le n^{-p/s} \le (k/n)^{p/s}$.

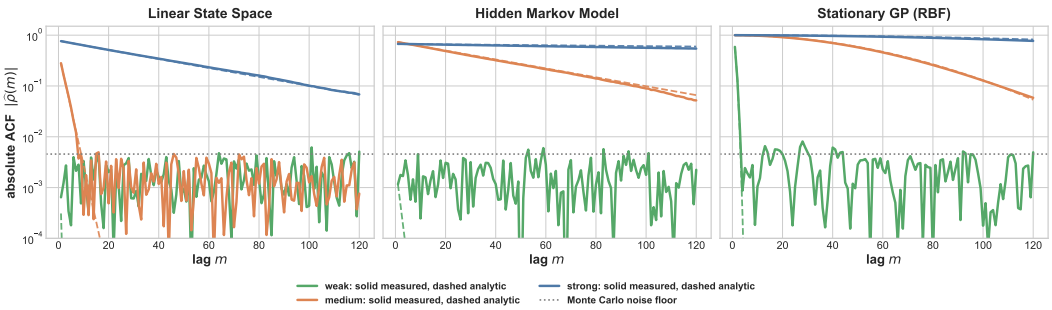


Figure 5: Assumption 1, correlation view. Mean absolute autocorrelation of the generated series ($d = 1$, $n = 6400$, 30 repetitions; solid) against the analytic decay rates (dashed; exact $\frac{6}{\pi}\arcsin(r_m/2)$ overlays for the Gaussian-copula models, geometric guide at the proved rate λ_2^m for the HMM). Measured autocorrelations track the corresponding analytic or model-implied decay curves until the Monte Carlo noise floor (dotted).

Step 3 (far tail via Hölder). Let $E_J := \{R_{n,k}(x) > 2^{J_0} r_*\}$ and $M_n := \max_{1\le i\le n}\|X_{n,i}-x\|$, so that $R_{n,k}(x) \le M_n$ almost surely and $\mathrm{E}[M_n^q] \le \sum_{i=1}^n \mathrm{E}\|X_{n,i}-x\|^q = nM_q$ by the common marginal. At $j = J_0$ one has $2^{J_0 s}k \asymp c_- r_0^s\, n/8$, so the exponential term in Step 1 is $e^{-cn/\log n} \le n^{-A}$ for all large n, whence $\mathrm{P}(E_J) \le Cn^{-A}$. Hölder's inequality with exponents $\big(q/p,\ (1-p/q)^{-1}\big)$ then gives

$$\begin{aligned}\mathrm{E}\big[R_{n,k}(x)^p\,\mathbf{1}_{E_J}\big] &\le \mathrm{E}[M_n^q]^{p/q}\,\mathrm{P}(E_J)^{1-p/q} \le C\,(nM_q)^{p/q}\,n^{-A(1-p/q)}\\ &= C'\,n^{p/q-A(1-p/q)} \le C'\,n^{-p/s},\end{aligned}$$

by the choice $A \ge A_0$, and $n^{-p/s} \le (k/n)^{p/s}$ since $k \ge 1$. Combining Steps 2–3 completes the proof. □

Remark 5 (Choice of the blocking exponent)**.** *Theorem 2 allows an arbitrary remainder $C_A n^{-A}$. For compact support, $A > p/s$ suffices; with only a finite qth marginal moment, the proof above uses $A \ge (p/s + p/q)/(1-p/q)$. The threshold on $k/\log n$ and the constants may therefore depend on the moment order, while the rate remains $(k/n)^{p/s}$.*

B.4 Empirical diagnostics for the synthetic assumptions

The preceding appendix verifies the assumptions analytically. Here we provide complementary empirical diagnostics using the same data-generating mechanisms, dependence-parameter settings, and global random seed (20260504) as Experiments 1 and 2. For Experiment 2, we also use the same fixed embedding matrix Q. The diagnostic sample sizes, replication counts, and evaluation grids are chosen specifically for these checks and are stated below. These diagnostics do not estimate the full past–future mixing coefficients. Rather, they check observable implications of the proved models, including autocorrelation decay, pairwise event-level dependence, effective sample size, and local ball-mass scaling. Wherever an exact finite-dimensional prediction is available, the corresponding overlay uses that analytic prediction with *no fitted parameters*. Table 4 collects the headline numbers.

Assumption 1, correlation view: decay at the proved rates For each Experiment-1 generator and dependence level ($d = 1$, $n = 6400$, 30 replications) we compute the mean sample autocorrelation of the series and overlay the analytic prediction. For the two Gaussian-copula models the overlay is exact: if the latent Gaussian pair at lag m has correlation r_m, then $\mathrm{Corr}\big(\Phi(w_t), \Phi(w_{t+m})\big) = \frac{6}{\pi}\arcsin(r_m/2)$, with $r_m = a^2\rho^m$ for the LSS and $r_m = e^{-m^2/(2\ell^2)}$ for the GP; for the HMM the guide is the proved latent rate λ_2^m of Proposition B.2, anchored at lag one. Figure 5 shows the result: the measured autocorrelations track the corresponding analytic or model-implied decay curves until they reach the Monte Carlo noise floor. This is consistent with geometric decay for the LSS and HMM and with faster-than-geometric pairwise correlation decay for the GP. The strong settings decay slowly, with characteristic scales of 50–1000 steps — the intended stress regime.

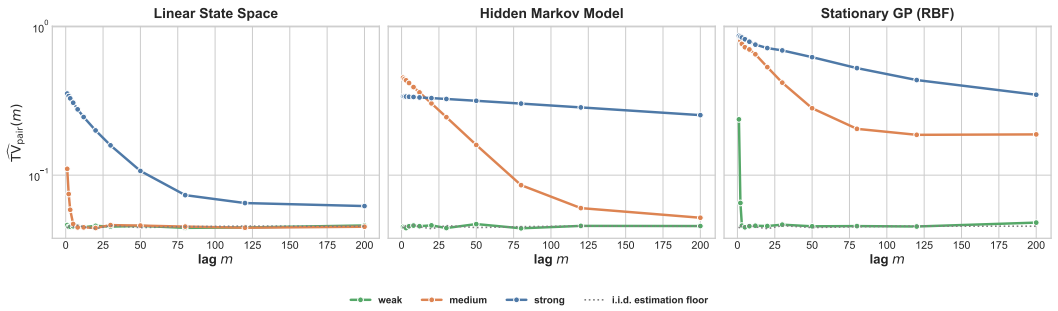


Figure 6: Assumption 1, event view. Pairwise total-variation dependence $\widehat{\mathrm{TV}}_{\mathrm{pair}}(m)$ between X_t and X_{t+m} (histogram estimator, 10×10 equiprobable bins; $d = 1$, $n = 6400$, 30 repetitions). The gray dotted curve is the i.i.d. estimation floor; reaching it is indistinguishable from independence. Pairwise event-level dependence decays toward the i.i.d. estimation floor with the expected ordering across dependence strengths.

Assumption 1, event view: a total-variation surrogate Autocorrelation measures only linear dependence, whereas Assumption 1 constrains dependence between *events*. As a distribution-level check we estimate the dependence between X_t and X_{t+m} in total variation,

$$\widehat{\mathrm{TV}}_{\mathrm{pair}}(m) \;=\; \tfrac{1}{2} \sum_{i,j=1}^{B} \left| \hat{P}_{ij}(m) - \hat{p}_i \hat{q}_j \right|,$$

where $\hat{P}(m)$ is the $B \times B$ histogram of the pairs (X_t, X_{t+m}) ($B = 10$ equiprobable bins, since the marginals are exactly uniform by Proposition B.4), in the spirit of histogram estimators for pairwise dependence such as [48]. This statistic measures the total-variation distance between the binned pair law of (X_t, X_{t+m}) and the corresponding product law. It is a pairwise event-level diagnostic, not an estimator of the full past–future β- or α-mixing coefficient. The i.i.d. generator provides the estimation floor of this estimator at this sample size: a curve that reaches the floor is indistinguishable from independence. Figure 6 shows that the pairwise event-level dependence decays toward the i.i.d. estimation floor, with the same qualitative ordering of weak, medium, and strong dependence as in Figure 5. This provides a nonlinear diagnostic complementary to autocorrelation.

From Assumption 1 to Figure 1: effective sample size The theory predicts that dependence enters the radius bounds only through γ in the constants and through the $k \gtrsim \log n$ regime — the exponent of $(k/n)^{1/s}$ is dependence-free. The natural finite-sample summary of Assumption 1 is the integrated autocorrelation time $\hat{\tau} = 1 + 2\sum_{m \geq 1} \widehat{\mathrm{acf}}(m)$ (Sokal windowing) and the effective sample size $n_{\mathrm{eff}} = n/\hat{\tau}$. Figure 7 plots the relative error of the fitted Figure 1 slopes (mean over p, $d = 1$) against n_{eff} at $n = 6400$. The pattern is sharp: every configuration with $n_{\mathrm{eff}} \gtrsim 2500$ reproduces the predicted slope p/d with the same $\approx 3\%$ error as the i.i.d. baseline, every deviating configuration has $n_{\mathrm{eff}} \leq 90$ (provable dependence horizons of 50–1000 steps against $n = 6400$), and the error generally increases as n_{eff} falls. The pattern is consistent with the interpretation that the deviations under strong dependence are finite *effective*-sample effects rather than a change in the asymptotic exponent.

Assumptions 2–3: local-mass exponent and constant We estimate $\mu(B(x, r))$ by the empirical fraction of sample points in $B(x, r)$, averaged over 200 interior query points $x \in [0.2, 0.8]^d$ and over repetitions ($d = 1$: $n = 20{,}000$, 20 repetitions; $d = 3$: $n = 40{,}000$, 8 repetitions), on a logarithmic radius grid $r \in [0.02, 0.15]$ chosen so that all balls stay inside $[0, 1]^d$. Two readouts follow. First, the fitted log–log slope $\hat{s}$ of $\hat{\mu}(B(x, r))$ against r estimates the local-mass exponent: across *every* model and dependence level, $\hat{s} \in [1.000, 1.001]$ for target $d = 1$ and $\hat{s} \in [2.976, 3.007]$ for target $d = 3$ (Table 4). Second, the ratio $\hat{\mu}(B(x, r))/(\omega_d r^d)$ in Figure 8 is approximately flat in r, as expected under the local scaling $\mu(B(x, r)) \asymp r^d$, at level 1 for the product-uniform models (Proposition B.4) and at the query-averaged copula density, also ≈ 1, for the LSS with $d = 3$ (Proposition B.5). This is consistent with the division of labor behind the experiments: dependence deforms the joint law

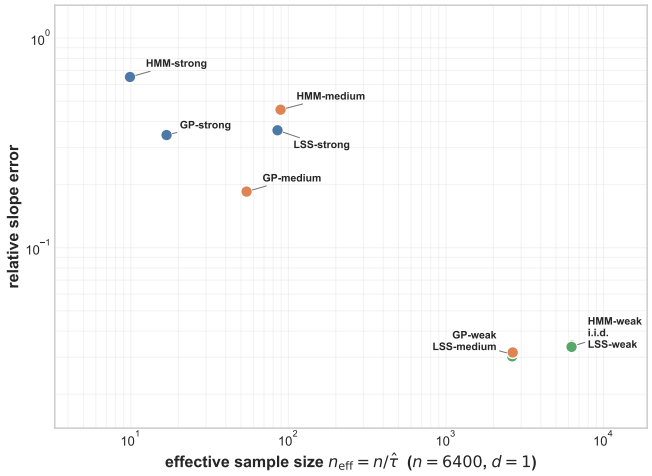


Figure 7: Relative error of the fitted slopes in Figure 1 (mean over p, $d = 1$, $n = 6400$) against the measured effective sample size $n_{\text{eff}} = n/\hat{\tau}$. Configurations with $n_{\text{eff}} \gtrsim 2500$ match the i.i.d. accuracy; all deviating configurations have $n_{\text{eff}} \leq 90$.

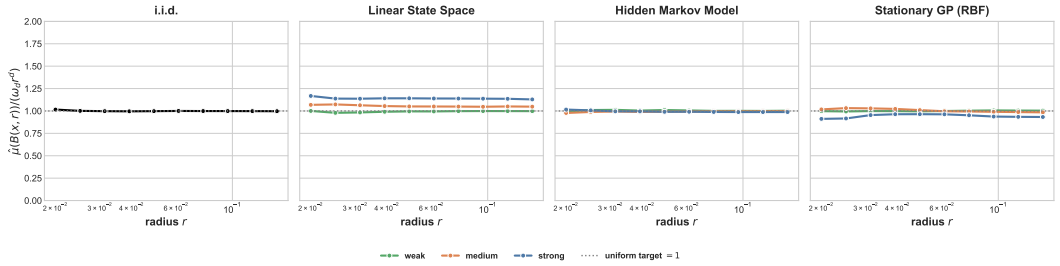


Figure 8: Assumptions 2–3. Ratio $\hat{\mu}(B(x,r))/(\omega_d r^d)$ for $d = 3$, averaged over 200 interior queries ($n = 40{,}000$ per repetition). Approximate flatness over the displayed radius range is consistent with $\mu(B(x,r)) \asymp r^d$; the level identifies the local constant (1 for the exactly uniform models; the query-averaged copula density for the LSS). The fitted exponents are $\hat{s} \in [2.976, 3.007]$ (Table 4).

across time (Assumption 1) but leaves the marginal local geometry, hence the exponent $s = d$ probed by the slope diagnostic, untouched at every dependence level.

Experiment 2: mixing and the intrinsic exponent The Experiment-2 diagnostics use the exact generator of Remark 4 (same embedding Q and seed as the experiment; $s \in \{3, 5\}$, $\rho \in \{0.3, 0.6\}$). For Assumption 1, every observed coordinate of $Y_t \in \mathbb{R}^{20}$ is a fixed linear function of the latent AR(1), so its autocorrelation is exactly ρ^m: the left panel of Figure 9 shows the measured curves on this parameter-free overlay down to the Monte Carlo noise floor, and the measured $\hat{\tau}$ is within 2% of the analytic $(1+\rho)/(1-\rho)$ (1.83 vs. 1.86 at $\rho = 0.3$; 4.07 vs. 4.00 at $\rho = 0.6$), giving $n_{\text{eff}} \approx 3500$ and 1570 at $n = 6400$ — consistent with moderate finite-sample dependence in Experiment 2. For Assumptions 2–3, we estimate the ball mass of the embedded cloud in $\mathbb{R}^{20}$ at on-support queries $y = Q(z^\top, 0, \ldots, 0)^\top$ with $z \sim N(0, I_s)$ ($n = 40{,}000$, 6 repetitions, 200 queries, $r \in [0.45, 1.1]$). Averaged over independent on-support queries, $Z - z \sim N(0, 2I_s)$, so the exact mass is $F_{\chi^2_s}(r^2/2)$ — an s-dimensional quantity, free of ρ and of any fitted parameter. The right panel of Figure 9 shows the measured mass lying *on* this curve: the fitted exponents are $\hat{s} \in [2.83, 2.85]$ (target $s = 3$) and $\hat{s} \in [4.81, 4.85]$ (target $s = 5$), against exact slopes of the analytic curve over the same radii of 2.84 and 4.81. (The exact slope sits slightly below s at finite radii because the Gaussian density varies within the ball; it tends to s as $r \to 0$, and the measured values match it to two decimals.) The local-mass exponent of the 20-dimensional data is therefore the *intrinsic* dimension s — an ambient exponent of 20 would fall by roughly twelve decades across this radius grid — which is precisely the mechanism that the PCA entropy experiment exploits. Assumption 4 fails for the Gaussian design but is not needed: Proposition B.6 applies with any $q > p$.

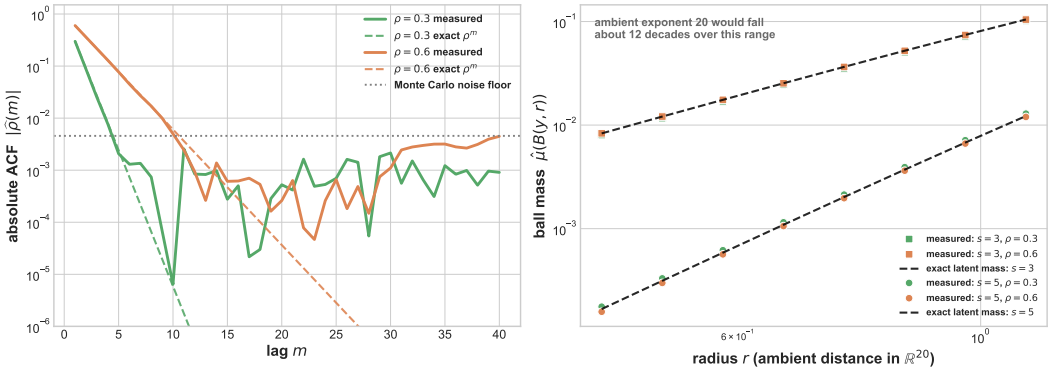


Figure 9: Experiment-2 diagnostics on the exact generator. Left: mean absolute autocorrelation of the observed coordinates in $\mathbb{R}^{20}$ against the exact ρ^m (dashed; no fitted parameters). Right: empirical ball mass of the embedded cloud at on-support queries against the exact s-dimensional latent mass $F_{\chi^2_s}(r^2/2)$ (dashed): the measured points lie on the s-dimensional curves, with fitted exponents $\hat{s} \approx 2.84$ and 4.81–4.85 for targets $s = 3$ and $s = 5$ — the intrinsic exponent, not the ambient dimension 20.

Model	Strength	Rate	γ	$\hat{\tau}$	n_{eff}	$\hat{s}$ (d=1)	$\hat{s}$ (d=3)
i.i.d.	—	—	∞	1.02	6302	1.000	2.995
LSS	weak	$\rho = 0.02$	3.91	1.02	6265	1.000	3.006
LSS	medium	$\rho = 0.60$	0.51	2.41	2652	1.001	2.989
LSS	strong	$\rho = 0.98$	0.020	75.1	85	1.000	2.991
HMM	weak	$\lambda_2 = 0.40$	0.92	1.02	6307	1.000	2.999
HMM	medium	$\lambda_2 = 0.98$	0.020	71.8	89	1.000	3.007
HMM	strong	$\lambda_2 = 0.999$	0.0010	648	9.9	1.000	2.988
GP–RBF	weak	$\ell = 1$	—	2.43	2633	1.000	3.005
GP–RBF	medium	$\ell = 50$	—	118	54	1.000	2.976
GP–RBF	strong	$\ell = 200$	—	380	17	1.001	3.006

Table 4: Assumption diagnostics for the Experiment-1 generators. "Rate" is the analytic per-step dependence rate (Propositions B.1 and B.2: $\alpha_m \le \frac{1}{4}\rho^m$ and $\alpha_m \le \frac{1}{4}\lambda_2^m$; for the GP, Proposition B.3 gives the super-geometric bound $\alpha_m \le \frac{1}{4}K_\ell e^{-m^2/(2\ell^2)}$, whose conservative constants are discussed in Remark 3, and the pairwise dependence horizon is ℓ), γ the implied exponent in Assumption 1, $\hat{\tau}$ and $n_{\text{eff}} = n/\hat{\tau}$ the measured integrated autocorrelation time and effective sample size at $n = 6400$ ($d = 1$), and $\hat{s}$ the fitted local-mass exponent of Assumptions 2–3 (targets $d = 1$ and $d = 3$).

Summary Table 4 summarizes the analytic rates alongside the empirical diagnostics computed on the exact implementation ($d = 1$, $n = 6400$, 40 repetitions): the integrated autocorrelation time $\hat{\tau}$, the effective sample size $n_{\text{eff}} = n/\hat{\tau}$, and the fitted local-mass exponents $\hat{s}$ (200 interior queries). The measured $\hat{\tau}$ matches the analytic dependence horizons, the configurations whose Figure 1 slopes deviate from p/d are exactly those with $n_{\text{eff}} \lesssim 10^2$ at $n = 6400$, and $\hat{s} \approx d$ throughout.

B.5 Experiment 3 (Section 4.3) implementation details

We follow the real-data benchmark setup of [22] for long-horizon forecasting, short-horizon forecasting, and classification, using the same task definitions, evaluation metrics, and preprocessing procedures. In addition to the source-to-target short-horizon setting considered in [22], we also report a within-group short-horizon evaluation. The setup for each task is as follows.

- **Long-horizon forecasting.** We report 8 datasets from the Informer long-horizon forecasting benchmark: ETTm1, ETTm2, ETTh1, ETTh2, Electricity, Traffic, Weather, and ILI. The evaluation metrics are MSE and mean absolute error (MAE). We use a look-back window length of 512 and a horizontal 60/10/30 train/validation/test split, with forecast horizons $\{24, 36, 48, 60\}$ for ILI and $\{96, 192, 336, 720\}$ otherwise. Hyperparameters are selected

by 5-fold time-series cross-validation when the training set contains at least 20 windows, and by validation-holdout tuning otherwise.

- **Short-horizon forecasting.** We use 6 benchmark groups from the Monash time series forecasting archive: M4-Yearly, M4-Quarterly, M4-Monthly, M3-Yearly, M3-Quarterly, and M3-Monthly. The evaluation metric is sMAPE, and the input length is 512. In the source-to-target transfer setting, hyperparameters are selected on a source group and evaluated directly on a different target group with the same frequency. We also report a within-group short-horizon evaluation, which is not included in the original MOMENT benchmark. In this setting, each dataset group is randomly split into a 70% source portion and a 30% target portion; hyperparameters are selected by 5-fold cross-validation on the source portion, and evaluation is performed on the target portion from the same group.
- **Classification.** We use 91 datasets from the UCR classification archive. The evaluation metric is accuracy. Inputs are represented as length-512 time series, and we use the official train/test splits provided by the UCR archive. Hyperparameters are selected using only the official training split, via stratified cross-validation when feasible, and final performance is evaluated on the official test split.

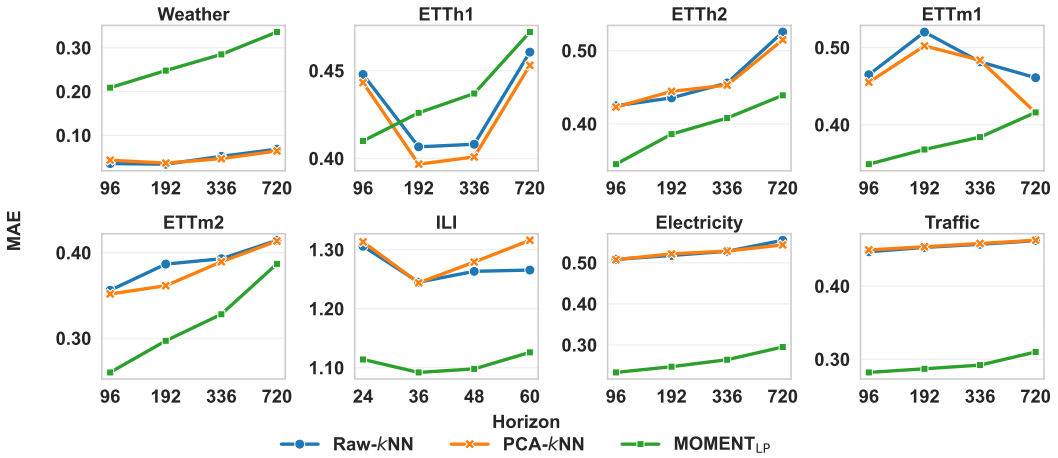


Figure 10: Comparison of long-horizon forecasting using test MAE across datasets and forecast horizons. It compares Raw-kNN and PCA-kNN with MOMENT across benchmark settings.

B.6 Real-data radius-scaling diagnostics

We complement the predictive evaluations in Section 4.3 with radius-scaling and local-mass diagnostics on representative real-data benchmarks. We use standardized input features for Raw-kNN and lower-dimensional PCA representations for PCA-kNN.

Diagnostic settings. For each dataset, we use at most 2000 reference samples and 300 query points. When these limits are exceeded, we sample rows uniformly without replacement and retain their original order. The short-horizon diagnostics use the source-to-target transfer setting, with M4 groups providing reference samples and the corresponding M3 groups providing query points. Standardization and PCA are fitted on the selected reference samples and applied to the query points. PCA uses up to 32 components, subject to the available sample size and feature dimension. We vary n using reference-sample prefixes corresponding to fractions $\{0.25, 0.4, 0.6, 0.8, 1.0\}$. For each eligible prefix, we combine neighbor counts $\{3, 5, 9, 15, 21, 31\}$ with rounded values of n^β for $\beta \in \{0.3, 0.5, 0.7\}$, retaining only counts smaller than n. For the local-mass diagnostic, we select radii at 12 equally spaced quantile levels between the 5th and 35th percentiles of the positive query-to-reference distances. We use 200 bootstrap replicates with a block size of 16 selected query rows, reduced when fewer query points are available. The random seed is 13.

Radius-scaling diagnostic. Using the notation from Section 2, we denote the k-NN radius around a query point x by $R_{n,k}(x) := \|\mathcal{X}^n_{(k)}(x) - x\|$, where n is the sample size and k is the number of

Table 5: Long-horizon forecasting results. We report test MSE and MAE for Raw-kNN, PCA-kNN, and MOMENT; MOMENT results are taken from Table 2 of [22].

Setting		Raw-kNN		PCA-kNN		$\text{MOMENT}_{\text{LP}}$	
Dataset	Horizon	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	0.023	0.036	0.121	0.043	0.154	0.209
	192	0.002	0.034	0.003	0.037	0.197	0.248
	336	0.055	0.052	0.160	0.047	0.246	0.285
	720	0.095	0.068	0.103	0.064	0.315	0.336
ETTh1	96	0.287	0.448	0.281	0.443	0.387	0.410
	192	0.237	0.407	0.227	0.397	0.410	0.426
	336	0.239	0.408	0.229	0.401	0.422	0.437
	720	0.305	0.461	0.294	0.453	0.454	0.472
ETTh2	96	0.289	0.425	0.286	0.423	0.288	0.345
	192	0.306	0.436	0.309	0.444	0.349	0.386
	336	0.332	0.456	0.318	0.453	0.369	0.408
	720	0.433	0.525	0.410	0.515	0.403	0.439
ETTm1	96	0.314	0.465	0.297	0.455	0.293	0.349
	192	0.382	0.520	0.356	0.502	0.326	0.368
	336	0.325	0.481	0.328	0.484	0.352	0.384
	720	0.304	0.461	0.247	0.416	0.405	0.416
ETTm2	96	0.212	0.356	0.206	0.352	0.170	0.260
	192	0.245	0.387	0.215	0.361	0.227	0.297
	336	0.252	0.393	0.246	0.389	0.275	0.328
	720	0.279	0.415	0.275	0.414	0.363	0.387
ILI	24	3.038	1.306	3.065	1.313	2.728	1.114
	36	2.723	1.245	2.728	1.244	2.669	1.092
	48	2.915	1.263	2.848	1.279	2.728	1.098
	60	2.872	1.265	2.931	1.316	2.883	1.126
ECL	96	0.476	0.508	0.470	0.508	0.136	0.233
	192	0.493	0.518	0.493	0.522	0.152	0.247
	336	0.509	0.528	0.502	0.529	0.167	0.264
	720	0.545	0.555	0.520	0.544	0.205	0.295
Traffic	96	0.396	0.447	0.392	0.449	0.391	0.282
	192	0.405	0.453	0.400	0.454	0.404	0.287
	336	0.408	0.457	0.405	0.458	0.414	0.292
	720	0.409	0.462	0.410	0.463	0.450	0.310

Table 6: Classification accuracy across 91 UCR datasets. We report summary statistics for Raw-kNN, PCA-kNN, and MOMENT; MOMENT results are taken from Table 4 of [22].

Statistic	Raw-kNN	PCA-kNN	MOMENT_0
Mean	0.748	0.748	0.794
Median	0.753	0.760	0.815
Std.	0.153	0.151	0.147

Table 7: Short-horizon forecasting results measured using sMAPE. We compare Raw-kNN and PCA-kNN transfer results with MOMENT results from Table 3 of [22] under the same source-to-target evaluation setting.

Setting		Raw-kNN	PCA-kNN	$\text{MOMENT}_{\text{LP}}$	
Dataset	Frequency			M4	FR
M3	Yearly	22.14	22.50	16.74	16.97
	Quarterly	17.67	17.24	10.09	10.62
	Monthly	21.05	21.77	16.04	16.90
M4	Yearly	36.48	34.91	–	14.84
	Quarterly	34.48	34.93	–	12.02
	Monthly	31.44	32.86	–	15.80

Table 8: Short-horizon forecasting results measured using sMAPE under the within-group evaluation setting. We compare Raw-kNN and PCA-kNN with MOMENT results from Table 3 of [22].

Target series		Raw-kNN	PCA-kNN	$\text{MOMENT}_{\text{LP}}$	
Dataset	Frequency			M4	FR
M3	Yearly	20.56	21.32	16.74	16.97
	Quarterly	14.33	14.38	10.09	10.62
	Monthly	15.78	16.17	16.04	16.90
M4	Yearly	16.65	16.69	–	14.84
	Quarterly	14.51	14.84	–	12.02
	Monthly	15.37	15.53	–	15.80

nearest neighbors. We compute this radius in either the standardized observation space or the PCA representation. We vary n and k, compute the average radius over query points, and fit its slope versus k/n on a log-log scale. This diagnostic examines how the empirical radius varies with k/n, motivated by the canonical scale $(k/n)^{1/s}$ under the theoretical assumptions, where s is the local geometric dimension.

Local-mass diagnostic. We also examine how empirical mass accumulates around the query points. For each radius r, we compute $\widehat{\mu}(B(x,r))$, the empirical fraction of reference sample points falling in the ball $B(x,r)$. We average this quantity over query points and fit the logarithm of the average empirical mass versus $\log r$. The resulting slope summarizes mass growth over the evaluated radius range.

Bootstrap intervals. We obtain 95% block-bootstrap intervals for the radius slopes by resampling blocks of consecutive rows in the selected query set. Within each replicate, we use the same resampled query indices across all (n,k) configurations and refit the slope. The reference samples and fitted representations remain fixed. These intervals summarize variability under query-block resampling; they do not include variability from resampling the reference data or refitting the representations. Consecutive selected query rows need not correspond to consecutive observations in the original time series.

Results. Tables 9, 10, and 11 report the fitted radius slopes, their 95% block-bootstrap intervals, and the local-mass slopes for long-horizon forecasting, short-horizon forecasting, and UCR classification, respectively. Across these benchmarks, the fitted radius slopes are positive for both Raw-kNN and PCA-kNN representations, with the reported intervals bounded away from zero. This indicates that the average empirical k-NN radius shrinks as k/n becomes smaller over the evaluated range, qualitatively consistent with the canonical scaling $(k/n)^{1/s}$. The PCA representation gives larger fitted radius slopes on these datasets. The local-mass fits also yield finite, positive empirical local exponents over the evaluated radius range. These diagnostics connect the real-data results to the radius-scaling and local-mass behavior studied in the theory, without establishing the exact exponent $1/s$ or verifying the full local-mass and α-mixing assumptions.

B.7 Dataset availability

We use two types of datasets in this paper: synthetic datasets and real-world time-series benchmark datasets. The synthetic datasets in Experiments 4.1 and 4.2 are generated from the data-generating processes described in Appendices B.1 and B.2, respectively, using the global random seed 20260504. The real-world datasets in Experiment 4.3 are taken from the Time Series PILE [22], which includes datasets from the Informer long-horizon forecasting benchmark, the Monash time series forecasting archive, and the UCR classification archive. These benchmark datasets are publicly available from their original sources.

B.8 Code availability

Code for synthetic data generation, preprocessing, hyperparameter tuning, evaluation, and figure reproduction will be made publicly available upon publication. Experiments 4.1 and 4.2 were run

Table 9: Radius-scaling and local-mass diagnostics for long-horizon forecasting. The 95% block-bootstrap intervals apply to the radius slopes.

Dataset	Representation	Radius slope	95% CI	Local-mass slope
Weather	Raw-kNN	0.176	[0.165, 0.185]	2.153
	PCA-kNN	0.237	[0.221, 0.254]	1.989
ETTh1	Raw-kNN	0.213	[0.204, 0.224]	2.601
	PCA-kNN	0.227	[0.214, 0.239]	2.441
ETTh2	Raw-kNN	0.158	[0.133, 0.186]	3.662
	PCA-kNN	0.183	[0.152, 0.217]	3.337
ETTm1	Raw-kNN	0.238	[0.223, 0.251]	2.139
	PCA-kNN	0.247	[0.230, 0.261]	2.093
ETTm2	Raw-kNN	0.199	[0.178, 0.221]	3.026
	PCA-kNN	0.201	[0.180, 0.223]	3.011
Electricity	Raw-kNN	0.087	[0.079, 0.095]	5.451
	PCA-kNN	0.108	[0.097, 0.119]	5.124

Table 10: Radius-scaling and local-mass diagnostics for short-horizon forecasting. The 95% block-bootstrap intervals apply to the radius slopes.

Dataset	Representation	Radius slope	95% CI	Local-mass slope
M3 Yearly	Raw-kNN	0.184	[0.167, 0.211]	2.462
	PCA-kNN	0.213	[0.197, 0.239]	2.394
M3 Quarterly	Raw-kNN	0.172	[0.162, 0.185]	2.815
	PCA-kNN	0.240	[0.224, 0.256]	2.617
M3 Monthly	Raw-kNN	0.145	[0.134, 0.160]	3.128
	PCA-kNN	0.206	[0.193, 0.225]	2.814

entirely on CPU on a cluster node with two AMD EPYC 9754 128-Core Processors, comprising 256 physical cores and 256 hardware threads. Reproducing these two experiments takes approximately 5 hours. Experiment 4.3 was run entirely on CPU on a Linux workstation with an AMD EPYC 9554 64-Core Processor, comprising 64 physical cores and 128 hardware threads, and takes approximately 4 hours to reproduce. No GPU acceleration was used.

Table 11: Radius-scaling and local-mass diagnostics for UCR classification. The 95% block-bootstrap intervals apply to the radius slopes.

Dataset	Representation	Radius slope	95% CI	Local-mass slope
Crop	Raw-kNN	0.212	[0.198, 0.232]	2.111
	PCA-kNN	0.214	[0.200, 0.235]	2.101
ElectricDevices	Raw-kNN	0.080	[0.061, 0.095]	8.944
	PCA-kNN	0.142	[0.127, 0.155]	4.619
ECG5000	Raw-kNN	0.194	[0.168, 0.219]	2.619
	PCA-kNN	0.211	[0.187, 0.237]	2.442

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See **Abstract** and **Our contributions** in Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: See Section 3.2 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper reports three experiments: two using synthetic data and one using publicly available real-world time-series benchmark datasets. It also provides the preprocessing, evaluation, and implementation details needed to reproduce the main results; see Experiments 4.1, 4.2, and 4.3, and Appendix B.7 and B.8.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: See Appendix B.7 and B.8.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Experiments 4.1, 4.2, and 4.3, and Appendix B.7 and B.8.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our experiment does not include a statistical significance check.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix B.8.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Appendix B.8.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.