Query Lower Bounds for Diffusion Sampling

Zhiyang Xun
UT Austin
zxun@cs.utexas.edu

Eric Price
UT Austin
ecprice@cs.utexas.edu

Abstract

Diffusion models generate samples by iteratively querying learned score estimates. A rapidly growing literature focuses on accelerating sampling by minimizing the number of score evaluations, yet the information-theoretic limits of such acceleration remain unclear.

In this work, we establish the first score query lower bounds for diffusion sampling. We prove that for d-dimensional distributions, given access to score estimates with polynomial accuracy $\varepsilon = d^{-O(1)}$ (in any L^p sense), any sampling algorithm requires $\widetilde{\Omega}(\sqrt{d})$ adaptive score queries. In particular, our proof shows that, within any polynomial total-query budget, successful sampling requires searching over $\widetilde{\Omega}(\sqrt{d})$ distinct noise levels, providing a formal explanation for why multiscale noise schedules are necessary in practice.

1 Introduction

Diffusion models have become a central paradigm in modern generative modeling, enabling major advances in tasks ranging from high-fidelity image synthesis to scientific computing [SDWMG15, HJA20, SSDK+21, ND21, KAAL22]. The key to their success lies in an iterative formulation: rather than generating samples directly in a single shot, these models progressively transform simple noise into structured data by repeatedly evaluating the *smoothed scores* of the distribution [SE19, SSDK+21]. By reducing the task of sampling from complex high-dimensional multimodal distributions to estimating the scores, this framework has dramatically improved generative modeling capabilities.

A central goal within this framework is improving sampling efficiency. Since each iteration typically requires evaluating a large neural network, the computational cost is dominated by the total number of such evaluations (queries). This has motivated a large body of work on reducing query complexity, including higher-order numerical solvers, optimized discretization schemes, and flow matching techniques that straighten generation trajectories [LZB+22, LZB+25, LRLZ22, ZC23, ZBR+23, KAAL22, LCBH+23, LGL23]. These advances have steadily reduced the number of queries required in both theory and practice.

This rapid progress raises a basic theoretical question: as we aim to reduce the number of score queries, what are the intrinsic information-theoretic limits of diffusion sampling in high dimension? When access to the target distribution is limited to querying smoothed score estimates, is there an unavoidable lower bound on the number of such queries? In this work, we ask:

Can we establish a query-complexity lower bound for score-based sampling?

Recent theory has made substantial progress on *upper* bounds [CCL+23b, LLT22]. In particular, for targets with bounded second moments, one can sample in $\widetilde{O}(d)$ iterations given access to L^2-accurate smoothed score estimates [BDBDD24]. At the same time, practical diffusion samplers often produce high-quality samples in only a handful of steps, far fewer than worst-case theoretical guarantees. This

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

gap makes lower bounds essential: is a polynomial dependence on d truly unavoidable, or could the query complexity be reduced to $\mathrm{polylog}(d)$, or even to $O(1)$?

Ruling out this possibility turns out to be very nontrivial. Information-theoretically, *smoothed* score estimates are very powerful in sampling, suggesting that such efficiency improvements might indeed be feasible:

1. Each query of a smoothed score returns a full vector in $\mathbb{R}^d$, providing $\Omega(d)$ bits of information. Thus, if each query can extract a near-optimal amount of information, the intrinsic difficulty of sampling might not necessarily grow with dimension.
2. Although scores are by definition local quantities, the smoothed scores can reveal global properties: Gaussian convolution can encode global distributional structure into local gradients, potentially enabling efficient navigation of high-dimensional space.
3. Even though L^2-accurate score estimates only guarantee accuracy in expectation rather than at any individual query point, large smoothing levels can mitigate this gap: they ensure that queries land within the typical set of the smoothed distribution, where the L^2 guarantee translates into high-probability pointwise accuracy.

Furthermore, recent results hint at a gap between limitations of specific discretizations and the potential power of smoothed-score access. For instance, [JZL25, GNZ25] show that even for a d-dimensional Gaussian, standard discretizations can require $\Omega(\sqrt{d})$ steps due to error accumulation. On the other hand, work on high-accuracy regimes has begun to explore ε-accurate sampling with $\mathrm{polylog}(1/\varepsilon)$ dependence [GCS26, CCDR26], raising the possibility of sampling with $\mathrm{polylog}(d)$ steps on product distributions. These considerations leave open whether there is a fundamental, information-theoretic barrier to few-step diffusion sampling in the worst case.

1.1 Our Results

In this work, we establish the first query-complexity lower bounds for diffusion sampling. Our main message is that, under standard distributional assumptions and polynomially accurate score estimates, even achieving TV error below 0.99 requires $\widetilde{\Omega}(\sqrt{d})$ score queries in the worst case.

To state our bounds formally, we first introduce the necessary notation. Let π be a target distribution on $\mathbb{R}^d$. For any noise level $\sigma > 0$ we write

$$\pi_\sigma = \pi * \mathcal{N}(0, \sigma^2 I_d), \qquad s_{\pi,\sigma}(x) = \nabla \log \pi_\sigma(x),$$

for the Gaussian-smoothed distribution and its score. We model diffusion samplers as follows:

Definition 1.1 (Diffusion Sampling)**.** *A diffusion sampling algorithm $\mathcal{A}$ accesses a target distribution π exclusively via adaptive queries to an oracle for smoothed score estimates $\widehat{s}_\sigma(x)$, outputting a sample $\widehat{X}$ whose law aims to approximate π.*

To guarantee convergence, assumptions on the class of target distributions and the quality of the score estimates are needed. We work under two *strong* conditions (i.e., favorable to the algorithm) that are standard in the diffusion-theory literature; therefore, our lower bounds apply to a broad range of algorithms.

First, we assume the target is a Gaussian smoothing of a bounded-support distribution.

Assumption 1 (Bounded plus noise)**.** *For constant $R > 0$ and constant $\gamma \in (0, R/2)$, there exists a distribution π_{pre} supported on $[-R, R]^d$ such that*

$$\pi = \pi_{\mathrm{pre}} * \mathcal{N}(0, \gamma^2 I_d).$$

Bounded-plus-noise is stronger than bounded moments or subgaussian tails, and it automatically implies $M_2 = \mathbb{E}[\|X\|_2^2] = O(d)$ as well as bounds on all higher moments. The additional Gaussian smoothing ensures that π is infinitely differentiable and mirrors the smoothed target distributions typically sampled by diffusion models (e.g., via early stopping). Furthermore, this smoothing makes Total Variation (TV) convergence guarantees possible for sampling algorithms.

Second, we assume the score estimate is polynomially L^p-accurate. The case $p = 2$ matches the standard assumption in diffusion theory analysis, as it aligns with the standard score matching

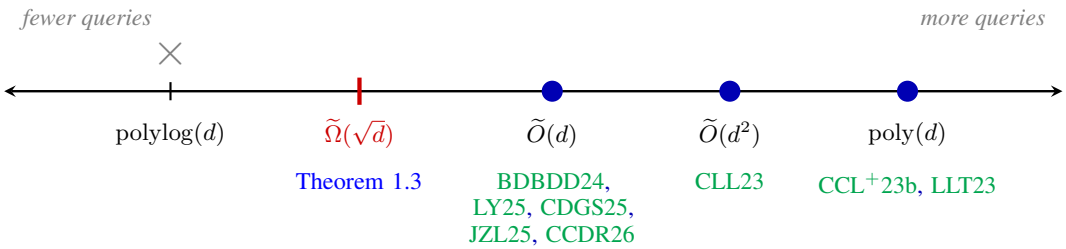


Figure 1: Score query complexity of diffusion sampling under Assumptions 1 and 2: representative upper bounds from prior work, alongside our $\widetilde{\Omega}(\sqrt{d})$ lower bound, which rules out $\mathrm{polylog}(d)$ query complexity in this setting.

training objective used in practice [Hyv05, Vin11, SE19, HJA20, SSDK+21, GPPX24]. We state the assumption for any constant $p \geq 2$ to cover a wider range of algorithms [TY24, XGP25].

Assumption 2 (L^p-accurate score)**.** *There exists a constant $p \geq 2$ and $\varepsilon_{\mathrm{err}} = 1/\operatorname{poly}(d)$ such that for all $\sigma > 0$,*

$$\mathop{\mathbb{E}}_{X \sim \pi_\sigma} \left[\|\widehat{s}_\sigma(X) - s_{\pi,\sigma}(X)\|_2^p \right] \leq \frac{\varepsilon_{\mathrm{err}}^p}{\sigma^p}.$$

Under Assumptions 1 and 2, the best known upper bound is that $\widetilde{O}(d)$ queries suffice to achieve small TV error. The bound can be achieved by analyzing the discretization error of standard algorithms like DDPM [BDBDD24, LY25, CDGS25]. We restate the result of [BDBDD24] below, translating their original KL bound to TV distance to facilitate comparison:

Theorem 1.2 ([BDBDD24])**.** *Under Assumptions 1 and 2, there exists a diffusion sampling algorithm (DDPM) that achieves TV error 0.01 using $\widetilde{O}(d)$ queries.*

Our main theorem shows that one cannot reduce the query complexity below $\widetilde{\Omega}(\sqrt{d})$ in the worst case, even when allowing TV error as large as 0.99; see Figure 1 for a comparison with existing upper bounds.

Theorem 1.3 (Main Theorem)**.** *Any diffusion sampling algorithm under Assumptions 1 and 2 with TV error less than 0.99 requires at least $\widetilde{\Omega}(\sqrt{d})$ queries.*

More recently, researchers have shown that beyond Assumptions 1 and 2, assuming additional Lipschitz regularity of the score can lead to algorithm acceleration, achieving sublinear iteration complexity in d [GCC25, LJ25]. For example, [ZHH+25] showed that if the score estimates are L-Lipschitz, then DDPM can require only $\widetilde{O}(L\sqrt{d})$ score queries; under their score-accuracy assumption, [JL25] gave an algorithm with $\widetilde{O}(\min\{d, L^{1/3}d^{2/3}, Ld^{1/3}\})$ queries, where L is the Lipschitz constant of the true scores.

Therefore, a useful perspective is that our hard instance has globally Lipschitz true smoothed scores and score estimates, with a Lipschitz parameter that scales linearly with dimension.

Remark 1.4 (Lipschitz scale of the smoothed score)**.** *In the hard instance underlying Theorem 1.3, for every $\sigma > 0$, both $s_{\pi,\sigma}$ and $\widehat{s}_\sigma$ are globally $O(d)$-Lipschitz.*

This implies that any algorithm with a query upper bound of $\widetilde{O}(L^a d^b)$ yields a complexity of $\widetilde{O}(d^{a+b})$ when instantiated on our hard instance. Theorem 1.3 therefore rules out query upper bounds of the form $\widetilde{O}(L^a d^b)$ with $a + b < 1/2$.

A more formal version of the theorem, which tracks the dependence on (R, γ), the oracle accuracy, and the explicit Lipschitz bound, is given in Theorem 2.1.

At the proof level, our lower bound isolates one intrinsic source of hardness: within any polynomial total-query budget, a successful diffusion sampler must scan through $\widetilde{\Omega}(\sqrt{d})$ distinct noise levels. This matches what diffusion algorithms do in practice, where scores are queried along a multiscale noise schedule, and provides a formal explanation for why such a design is necessary. Our proof proceeds by reduction to a hypothesis-testing problem: we construct a null distribution π_{null} and

a class of planted distributions $\mathcal{D}$, and show that $\widetilde{\Omega}(\sqrt{d})$ queries at distinct noise levels are needed to distinguish whether the score estimates come from π_{null} or from some π drawn from $\mathcal{D}$. In fact, with fewer queries the algorithm's execution on a random planted instance is, with high probability, identical to its execution on the null instance, so its output law is nearly a fixed distribution, which cannot approximate a random planted target. Section 2.2 outlines the proof.

Under Assumptions 1 and 2, a gap remains between our $\widetilde{\Omega}(\sqrt{d})$ lower bound and the best known $\widetilde{O}(d)$ upper bound. Closing this gap is an important open problem, and progress could come from either direction: stronger lower bounds that go beyond hypothesis testing to directly exploit the difficulty of producing a sample, or faster algorithms that leverage the structure of the noise-level scanning bottleneck identified here.

Wasserstein Distance. Another line of work analyzes diffusion samplers in Wasserstein distance. [AVD25] show that $\widetilde{O}(\sqrt{d}/\varepsilon)$ queries to a uniformly accurate score suffice for W_2 error ε under log-concavity-type conditions, and [GSO25] obtain W_2 guarantees under weak log-concavity; both bounds depend exponentially on parameters quantifying non-log-concavity. The same hard instances give a Wasserstein analog of Theorem 1.3.

Theorem 1.5 (Wasserstein Lower Bound)**.** *There is a constant $c > 0$ such that, for every $q \geq 1$, any diffusion sampling algorithm under Assumptions 1 and 2 with W_q error less than $c\sqrt{d}$ requires at least $\widetilde{\Omega}(\sqrt{d})$ queries.*

Outputting the origin, with no queries, already achieves W_q error $O(R\sqrt{d})$ for every such π and fixed q, so Theorem 1.5 says that, with fewer queries, no algorithm improves on this trivial output by more than a constant factor.

Subexponential Error Tail. While L^p accuracy is the standard assumption, it is also instructive to consider the impact of stronger error tail conditions. Recent work has shown that stronger error tail requirements might yield accelerated algorithms [GCS26].

Assumption 3 (Subexponential Error)**.** *For all $\sigma > 0$ and all $z \geq 0$, the model can query $\widehat{s}_\sigma(x)$ with the guarantee that*

$$\Pr_{X \sim \pi_\sigma}\left[\|\widehat{s}_\sigma(X) - s_{\pi,\sigma}(X)\|_2 \geq z\right] \;\leq\; 2\exp\left(-\frac{z\sigma}{\varepsilon_{\text{err}}}\right).$$

We show that, under a constant subexponential error model, $\Omega(d^{1/4})$ queries are still necessary.

Theorem 1.6 (Subexponential Tail)**.** *For any constant $\varepsilon_{\text{err}} > 0$, any algorithm that solves diffusion sampling under Assumptions 1 and 3 with TV error less than 0.99 requires at least $\Omega(d^{1/4})$ queries.*

Applicability to practical algorithms. Our oracle model is intended to capture the inference-time information available to a broad class of practical diffusion and diffusion-adjacent samplers: repeated evaluations of a learned network at user-chosen states and noise levels, without any additional access to the target distribution π. Appendix H explains how common network parameterizations, SDE/ODE-based samplers, and flow matching fit this model, and what it does not cover.

2 Proof Overview

2.1 Intuition on Hardness

To build intuition, we consider a simple hard family: mixtures of n well-separated Gaussians. Concretely, pick a uniformly random set of centers $S \subseteq \{-1, +1\}^d$ with $|S| = n$, and define

$$\pi_{\text{pre}} = \text{Unif}(S), \qquad \pi = \pi_{\text{pre}} * \mathcal{N}(0, 0.1 I_d).$$

Clearly, any nontrivial sampler must locate at least one center in S using score queries. Our proof shows that this is only possible when the algorithm queries a *narrow window* of smoothing levels determined by n. Since n is hidden and may range over $2^{\Theta(d)}$ possibilities, the sampler is forced to scan many different noise levels.

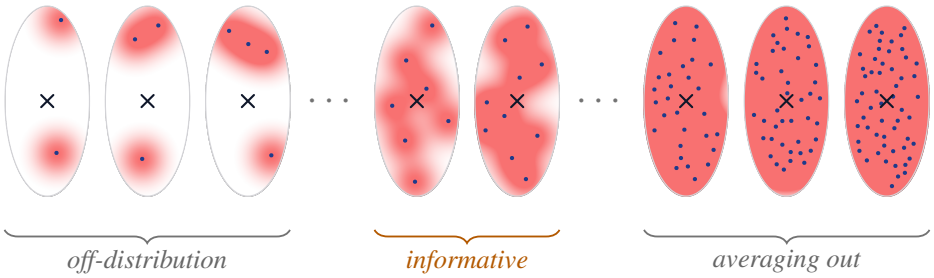


Figure 2: *Three regimes of the smoothed-score query at fixed τ.* With high probability, only a narrow range of $|S| = n$ produces an informative query at $\times$.

Throughout this subsection, we write τ for the total noise level, setting $\pi_\tau := \pi_{\text{pre}} * \mathcal{N}(0, \tau^2 I_d)$ with score $s_{\pi,\tau}$, and write s_τ^{null} for the null score obtained by replacing S with $\{-1,+1\}^d$.

By Tweedie's formula, for $X = Y + Z$ with $Y \sim \pi_{\text{pre}}$, $Z \sim \mathcal{N}(0, \tau^2 I_d)$, so that $X \sim \pi_\tau$, the score can be rewritten as

$$s_{\pi,\tau}(x) = \frac{m(x) - x}{\tau^2},$$

where $m(x) = \mathbb{E}[Y \mid X = x]$. Since Y is supported on S, $m(x)$ is a posterior average over the modes $y \in S$, with weights proportional to $\exp(-\|x-y\|_2^2/(2\tau^2))$. Thus a score query is informative only insofar as this posterior average *depends on the planted set* S.

As a reference, recall that s_τ^{null} is the score one would obtain if the adversary pretended $S = \{-1,+1\}^d$ (so the answer carries no information about S). We will see that for most τ, an adversary can answer essentially with s_τ^{null} while still satisfying the polynomial $L^p(\pi_\tau)$ accuracy guarantee, forcing the sampler to scan across τ.

Warmup: Gaussian as a sphere. A Gaussian in $\mathbb{R}^d$ concentrates on a thin shell: $\|Z\|_2 \approx \tau\sqrt{d}$. Imagine that it were *exactly uniform* on the sphere of radius $\tau\sqrt{d}$. Then for fixed (x, τ), $m(x)$ would simply average the sampled centers lying in the corresponding sphere around x. Let $N_{x,\tau}$ denote the expected number of points of S falling in this sphere.

This yields two uninformative extremes:

- If $N_{x,\tau} \gg \text{poly}(d)$, with high probability the averaging washes out the instance: by concentration, $m(x)$ becomes $1/\text{poly}(d)$-close to the null answer, so an adversary can answer with s_τ^{null}.
- If $N_{x,\tau} \ll 1/\text{poly}(d)$, then with high probability over S there are no points in the sphere. This implies that $\pi_\tau(x)$ has tiny mass, and the $L^p(\pi_\tau)$ guarantee does not constrain the oracle on these points. The adversary can again answer with s_τ^{null}.

Either way, a query at this τ reveals essentially nothing except that the chosen smoothing level is "wrong." The mechanism is illustrated in Figure 2.

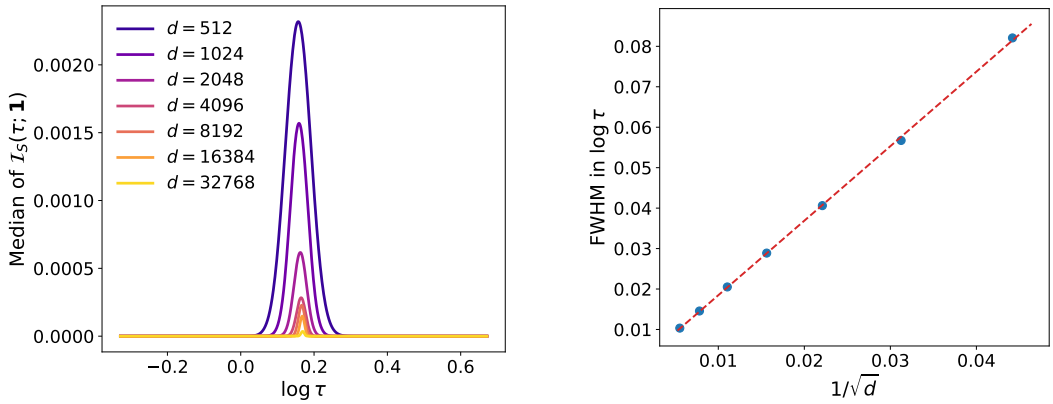
To get more information, τ must be tuned so that $N_{x,\tau}$ lies within an intermediate window of a $\text{poly}(d)$ range. Note that $N_{x,\tau}$ scales linearly with the underlying n. Therefore, each query would only rule out the range of n up to a $\text{poly}(d)$ factor, and there are

$$\log_{\text{poly}(d)}(2^{\Theta(d)}) = \Theta(d/\log d)$$

such ranges to check. Therefore, we need to query $\Theta(d/\log d)$ distinct smoothing levels to get one "not null" answer that actually helps sampling.

The real Gaussian gives $\sqrt{d}$. The analysis above fails in one place: a Gaussian is *not* uniform even on its thin shell. With $1 - 1/\text{poly}(d)$ probability,

$$\|Z\|_2^2 = \tau^2 d \pm \widetilde{O}(\tau^2\sqrt{d}),$$



(a) Median of the shell-resolved proxy for $\mathcal{I}_S(\tau;\, \mathbf{1})$, for $d = 512$ to 32768.

(b) FWHM versus $1/\sqrt{d}$.

Figure 3: **Informative window on the hypercube warm-up family** ($\kappa = 0.2$). The proxy concentrates near a single smoothing scale and narrows as d grows; the width scales as $1/\sqrt{d}$, consistent with the predicted $d^{-1/2}$ rate. See Appendix I for the precise definition of the proxy.

so the posterior is influenced by a whole band of squared distances of width $\Theta(\tau^2\sqrt{d})$, and the mass outside is negligible.

Pick two squared radii $r_\pm^2 = \tau^2 d \pm \Theta(\tau^2\sqrt{d})$ within this typical band. The per-center likelihood weights satisfy

$$\frac{\exp(-r_-^2/(2\tau^2))}{\exp(-r_+^2/(2\tau^2))} = \exp\left(\Theta\left(\frac{r_+^2 - r_-^2}{\tau^2}\right)\right) = \exp(\Theta(\sqrt{d})).$$

At the same time, the size of the corresponding Hamming shell at radius r_+ versus r_- differs by $\exp(\Theta(\sqrt{d}))$ as well.

When S is a random subset of size n, this compensation creates a gap between the two thresholds in the counting story: as n grows, one first exits the "empty" regime by hitting the high-surface-area part of the band, but only after an additional $\exp(\widetilde{\Theta}(\sqrt{d}))$ factor in n does one reach the true "averaging" regime, where even the low-surface-area (but high-weight) part is well populated. Equivalently, for fixed τ the informative window can pin down n only up to an $\exp(\widetilde{\Theta}(\sqrt{d}))$ multiplicative factor.

Therefore we must scan

$$\log_{\exp(\widetilde{\Theta}(\sqrt{d}))}\big(\exp(\Theta(d))\big) = \widetilde{\Theta}(\sqrt{d})$$

distinct smoothing levels to reliably hit an informative scale, yielding the $\widetilde{\Omega}(\sqrt{d})$ lower bound.

Synthetic illustration. Figure 3 visualizes the informative-window phenomenon on the hypercube warm-up family described above. For a random codebook S drawn from a random-coding ensemble with expected size $e^{\kappa d}$ on $\{-1,+1\}^d$ (with $\kappa = 0.2$), consider the local signal at the fixed query point $x = \mathbf{1}$:

$$\mathcal{I}_S(\tau;\, \mathbf{1}) \;:=\; \frac{\pi_\tau(\mathbf{1})}{\pi_\tau^{\mathrm{null}}(\mathbf{1})} \;\cdot\; \frac{1}{d}\left\|s_{\pi,\tau}(\mathbf{1}) - s_\tau^{\mathrm{null}}(\mathbf{1})\right\|_2^2,$$

where π_τ is the smoothed density of the mixture defined by S and π_τ^{null} is the smoothed null density (corresponding to $S = \{-1,+1\}^d$), with $s_{\pi,\tau}$ and s_τ^{null} their respective scores. This measures the per-coordinate Fisher signal at $x = \mathbf{1}$ between the planted and null distributions at noise level τ. When $\mathcal{I}_S(\tau;\, \mathbf{1})$ is negligible, a query at noise level τ carries little local planted-versus-null score signal at this point.

Figure 3a plots a tractable proxy for $\mathcal{I}_S(\tau;\, \mathbf{1})$, obtained by conditioning on shell occupancies (see Appendix I for details). The proxy concentrates around a single matching scale and becomes progressively narrower as d grows. Figure 3b shows that the full width at half maximum scales as

$1/\sqrt{d}$, consistent with the $d^{-1/2}$ prediction from the shell fluctuation argument above. This narrowing illustrates the mechanism behind the distinct-level lower bound.

2.2 Proof Roadmap

We now state the formal version of Theorem 1.3 and outline its proof.

Theorem 2.1. *For every $p > 0$, $\rho \in (0, 1/4)$, and constant $a \geq 1$, there exists $c = c(p, \rho, a) > 0$ such that the following holds. For every $R > 0$ and $\gamma \in (0, R/2)$, there exists $d_0 = d_0(p, \rho, a, R/\gamma)$ such that, for every integer $d \geq d_0$ and every $\varepsilon_{\mathrm{err}} \in (0, 1]$, there exists a distribution $\mathcal{D}$ over pairs $(\pi, \{\widehat{s}_\sigma\}_{\sigma>0})$ on $\mathbb{R}^d$ satisfying:*

(1) $\pi = \pi_{\mathrm{pre}} * \mathcal{N}(0, \gamma^2 I_d)$ for some π_{pre} supported on $[-R, R]^d$.

(2) For every $\sigma > 0$,

$$\mathop{\mathbb{E}}_{X\sim\pi_\sigma}\left[\|\widehat{s}_\sigma(X) - s_{\pi,\sigma}(X)\|_2^p\right] \leq \frac{\varepsilon_{\mathrm{err}}^p}{\sigma^p}.$$

(3) For every $\sigma > 0$, both $s_{\pi,\sigma}$ and $\widehat{s}_\sigma$ are globally L_σ-Lipschitz with

$$L_\sigma \leq \frac{1}{\gamma^2 + \sigma^2} + \frac{5R^2 d}{(\gamma^2 + \sigma^2)^2}.$$

(4) Writing

$$H := \log(d/\varepsilon_{\mathrm{err}}) + \log(R/\gamma), \qquad B_d := \frac{d \log(R/\gamma)}{\sqrt{dH} + H},$$

every adaptive algorithm that, on every execution path, makes at most d^a score queries in total and uses at most cB_d distinct smoothing levels (distinct values of σ) satisfies

$$\Pr_{(\pi,\{\widehat{s}_\sigma\})\sim\mathcal{D}}\left[\mathrm{d_{TV}}\left(\widehat{X}, \pi\right) \geq 1 - \rho\right] \geq 1 - \rho.$$

In particular, the same conclusion holds for every adaptive algorithm making at most cB_d score queries in total.

Appendices A–D prove Theorem 2.1; Appendix E extends it to non-uniform codebook weights of at most $\mathrm{poly}(d)/n$ each.

Given the intuition in Section 2.1, formalizing it into a real query-complexity lower bound still needs to handle several technical difficulties. The main issue is that the intuition just shows, for *each* x, that the informative region of σ can be small; the proof needs to construct explicit valid (Lipschitz, accurate) score functions for each σ, then show that every fixed x is uninformative over this family. Specifically:

(i) how to improve the bound by a factor of $\log(R/\gamma)$;

(ii) how to formally construct a score oracle that is L^p-accurate and Lipschitz, yet hides the information;

(iii) how to prove that each total noise level τ ($\tau^2 = \gamma^2 + \sigma^2$) gives information about S only for a roughly $\sqrt{d}$-wide range of $\log n$.

In Appendix B, we address (i) by using the product of $d/2$ blocks as the hard family, where each block contains $\Theta(R/\gamma)$ well-separated points on a circle of radius R, and take the codebook $S = (Y_1, \ldots, Y_n)$ to be n i.i.d. uniform points of this product. The admissible range of $\log n$ then has length $\Theta(d \log(R/\gamma))$, versus $\Theta(d)$ on the hypercube, without losing the Bernstein-type concentration analysis.

Appendix B also constructs the score estimate of (ii) by formalizing the intuition of Section 2.1: when the true score is already close in L^p to the null score, just use the null score; if not, answer with the true score where the codebook mixture's likelihood ratio against the null is large, and smoothly interpolate to the null score elsewhere. The concentration bound below makes this L^p-accurate for every codebook (Lemmas B.3 and B.4), and the smoothing, together with the natural smoothness of the true scores, gives the global Lipschitz bound (Lemma B.4).

Most of the effort goes into proving (iii), which is Proposition C.2. We sketch the proof below.

Dense side. Write $f_{S,\tau}$ and $f_{\mathrm{null},\tau}$ for the densities of the codebook mixture and of the null mixture (uniform over the full product support) at total noise τ, the analogs of π_τ and its null counterpart in Section 2.1, and define

$$\imath_\tau(y,x) := \log \frac{\varphi_{\tau^2}(x-y)}{f_{\mathrm{null},\tau}(x)},$$

with φ_{τ^2} the density of $\mathcal{N}(0,\tau^2 I_d)$. Write $\imath_\tau := \imath_\tau(Y,X)$ for Y uniform on the full product support and $X = Y + \tau Z$ with independent $Z \sim \mathcal{N}(0, I_d)$, so $I_\tau := \mathbb{E}[\imath_\tau]$ is the mutual information between Y and X.

Consider sampling X from the codebook mixture. By symmetry, we may assume that X is generated by the first codeword, $X = Y_1 + \tau Z$ with $Z \sim \mathcal{N}(0, I_d)$ independent of S. Then

$$\frac{f_{S,\tau}(X)}{f_{\mathrm{null},\tau}(X)} = \frac{1}{n}\sum_{j=1}^{n} e^{\imath_\tau(Y_j,X)}.$$

Every term with $j \neq 1$ has $\mathbb{E}[e^{\imath_\tau(Y_j,X)} \mid X] = 1$, as Y_j is uniform and independent of X, so the only bias away from null is the correlated term $e^{\imath_\tau(Y_1,X)}/n = e^{\imath_\tau(Y_1,X)-\log n}$, which is exponentially small once $\log n \gg \imath_\tau(Y_1,X)$. Therefore, once $\log n$ is much larger than I_τ, even at the codebook mixture's own sample points, the mixture has similar density to the null distribution and the signal cannot be seen. Heuristically, ne^{-I_τ} plays the role of $N_{x,\tau}$ in Section 2.1.

To make this quantitative, we need the concentration of $\imath_\tau$. By the product structure, $\imath_\tau = \sum_{k=1}^{d/2} \imath_\tau^{(k)}$ with the block terms i.i.d., and the proof of Lemma B.2 shows

$$\Pr\left[|\imath_\tau^{(k)} - \mathbb{E}\,\imath_\tau^{(k)}| \geq t\right] \leq 2e^{-t/C}$$

for an absolute constant C, independent of R/γ and of τ. Bernstein's inequality then gives $\imath_\tau = I_\tau \pm \widetilde{O}(\sqrt{d})$ with probability $1 - 1/\operatorname{poly}(d)$.

To translate the closeness in density at typical points to the L^p score closeness we want, we use Tweedie's formula and truncate the likelihood weights: once $\log n$ exceeds $I_\tau + \widetilde{O}(\sqrt{d})$, the true and null scores are close in L^2 (Lemma C.1), hence in L^p since their difference is bounded, so above this threshold the oracle is all-null with high probability.

Sparse side. For small n the signal is strong but restricted to the codewords' likelihood neighborhoods, which a blind query misses. By the same concentration bound, we can switch on the true score only where $f_{S,\tau}/f_{\mathrm{null},\tau} \geq e^{I_\tau - \widetilde{O}(\sqrt{d})}/n$ without hurting accuracy, since a sample from the codebook mixture exceeds this level through its own codeword's term (Lemma B.3); at a blind query point the likelihood ratio has mean one over the random codebook, so by Markov's inequality the query crosses the threshold with probability at most $ne^{-I_\tau + \widetilde{O}(\sqrt{d})}$. The oracle therefore stays null at any blind query until $\log n$ reaches $I_\tau - \widetilde{O}(\sqrt{d})$.

By putting the above two together, we can resolve (iii) and show that at every noise level, every blind query is answered with the null score with high probability, unless $\log n = I_\tau \pm \widetilde{O}(\sqrt{d})$ (Proposition C.2).

In Appendix D, we draw $\log n$ nearly uniformly from a range of length $\Theta(d\log(R/\gamma))$, so each noise level is informative with probability only $\widetilde{O}(1/(\sqrt{d}\log(R/\gamma)))$. A standard coupling argument then shows that, unless the algorithm uses $\widetilde{\Omega}(\sqrt{d}\log(R/\gamma))$ distinct levels, it behaves as if it faced the null oracle (Proposition D.1), so its output cannot land on the mass of the random planted target (Lemma D.2). This proves Theorem 2.1; the same witness, rescaled to be 1-Lipschitz, gives Theorem 1.5 (Corollary G.1).

All of the bounds above are for the L^p model, where the $1/\operatorname{poly}(d)$ oracle accuracy enters the window only through an $O(\log d)$ factor. For Theorem 1.6, the oracle must instead be subexponentially accurate, and this exponentially small requirement has logarithm $\Theta(\sqrt{d})$; the window widens from $O(\sqrt{d\log d})$ to $O(\sqrt{d\cdot\sqrt{d}})$, and the bound drops to $\Omega(d^{1/4})$ (Appendix F).

3 Related Work

While a rapidly growing literature studies the iteration complexity of diffusion sampling under various assumptions [CCL+23b, DB22, CDGS25, LLT22, CCL+23a, LLT23, CLL23, CDD23, LWCC24, LHE+24, WCW24, GZ25], lower bounds remain much less developed than convergence analyses. We discuss two directions that are closest in spirit: iteration lower bounds for specific diffusion algorithms and query lower bounds in the classical sampling literature.

3.1 Iteration Lower Bounds for Diffusion Models

Existing algorithmic lower bounds typically fix a discretization or a noise schedule. [JZL25] show that under a standard DDPM discretization schedule, even with exact scores and a Gaussian target, the KL error of the terminal marginal satisfies $\mathrm{KL} \gtrsim d/T^2$, which forces $T = \Omega(\sqrt{d}/\varepsilon)$ steps to achieve $\mathrm{KL} \le \varepsilon^2$. Related Gaussian lower bounds in Wasserstein metrics appear in [GNZ25]. These results identify barriers for specific solvers, but their analyses are tied to particular discretizations of specific stochastic processes and therefore cannot be generalized to the algorithm-agnostic lower bounds we seek.

It is also worth mentioning a potential route toward dimension-dependent query lower bounds via tensorization. For a product target $p^{\otimes d}$, KL divergence is additive:

$$\mathrm{D}_{\mathrm{KL}}(p^{\otimes d} \| q^{\otimes d}) = d \cdot \mathrm{D}_{\mathrm{KL}}(p \| q),$$

so achieving ε-KL accuracy on $p^{\otimes d}$ requires (ε/d)-KL accuracy on each marginal [BDBDD24]. If one-dimensional score-based sampling required $\mathrm{poly}(1/\varepsilon)$ queries, this could yield dimension-dependent lower bounds even for product targets. However, recent results [CCDR26] achieve $\mathrm{polylog}(1/\varepsilon)$ query complexity in one dimension under L^2-accurate score estimates, ruling out polynomial one-dimensional lower bounds in this oracle model.

3.2 Query Lower Bounds for Sampling

A classical oracle model writes the target as $\pi(x) \propto e^{-V(x)}$ and grants pointwise access to $V(x)$ and possibly its derivatives [CGL+22]. With first-order access, querying $\nabla V(x)$ is equivalent to querying the exact unsmoothed score

$$s_0(x) := \nabla \log \pi(x) = -\nabla V(x).$$

In the strongly log-concave and smooth setting, [CdDPL+24] prove dimension-dependent lower bounds that already appear for Gaussian targets. In particular, for a centered Gaussian with condition number κ, any algorithm with first-order oracle access needs

$$\widetilde{\Omega}\big(\min\{\sqrt{\kappa}\log d,\, d\}\big)$$

queries to achieve constant accuracy, matching block-Krylov type methods up to logarithmic factors. Sharp lower bounds are also known in one dimension [CGL+22].

For general non-log-concavity, it is known that $\exp(\Omega(d))$ queries are needed. A typical construction hides probability mass in one of exponentially many well-separated regions of $\mathbb{R}^d$, so that local values and gradients are essentially indistinguishable unless the algorithm queries near the right region [GLR18, HZ25, CGLL23].

Diffusion-style access changes the picture because the oracle is smoothed. Querying $s_{\pi,\sigma} = \nabla \log(\pi * \mathcal{N}(0, \sigma^2 I))$ aggregates information through Gaussian smoothing at scale σ. Large σ can reveal coarse global structure and rule out many candidate regions at once, while small σ refines local details. Therefore, our lower bounds require a different approach, and we show that the bottleneck is identifying an informative smoothing scale σ.

4 Conclusion

In this work, we establish the first information-theoretic lower bounds for diffusion sampling. We prove that for d-dimensional distributions, acquiring a nontrivial sample using L^p-accurate score

estimates requires $\widetilde{\Omega}(\sqrt{d})$ adaptive score queries, even under favorable assumptions of smoothness and polynomial estimation accuracy. These results show that the dependence of the computational cost on dimension is a fundamental barrier intrinsic to the score-based sampling paradigm, rather than a limitation of current solvers.

At a technical level, our proof shows that, within any polynomial total-query budget, successful sampling requires querying $\widetilde{\Omega}(\sqrt{d})$ distinct noise levels, providing an information-theoretic rationale for the multiscale noise schedules used in practice. The key mechanism is that the score signal is localized within a narrow window of noise levels; outside this window, oracle responses carry essentially no information about the target. This means that the bottleneck for few-step sampling is not discretization error but the number of noise levels queried.

Our work raises interesting open directions for future research:

1. **Tightening the Bound.** There remains a gap between our $\widetilde{\Omega}(\sqrt{d})$ lower bound and the best known $\widetilde{O}(d)$ upper bounds. Determining the precise query complexity is a major open question. On the lower-bound side, our reduction to hypothesis testing suggests a natural avenue: directly lower-bounding the difficulty of producing a sample, rather than extracting a single bit. On the upper-bound side, the noise-level scanning bottleneck identified here may inform the design of faster algorithms.
2. **Beyond the Smoothed-Score Oracle.** Our lower bounds apply when inference-time access to the target is only through repeated evaluations of a Gaussian-smoothed score, or an explicitly equivalent field as in rectified flow. A natural next step is to extend lower bounds to paradigms that change this primitive, including flow matching with non-Gaussian priors or alternative couplings, and distilled generators that produce samples in very few steps.
3. **Stronger Oracle and Instance Constraints.** Our construction is adversarial within Assumptions 1 and 2, with no restriction on how the score is realized. A natural refinement is to require $\widehat{s}_\sigma$, or even $s_{\pi,\sigma}$ itself, to lie in the function class of polynomial-sized neural networks, or to grant access to the *exact* smoothed score in place of an L^p-accurate estimate. Both restrictions remove substantial degrees of freedom in our construction and whether the $\widetilde{\Omega}(\sqrt{d})$ barrier persists under such constraints is open.
4. **Real-World Structure.** The empirical success of few-step sampling suggests that real-world data possess structures stronger than generic smoothness. Identifying specific geometric properties, such as low intrinsic dimension, that enable practical algorithms to break the worst-case $\sqrt{d}$ barrier is a promising direction for bridging theory and practice.

Acknowledgments

ZX thanks Chihao Zhang for a helpful discussion on query lower bounds for log-concave sampling. This work is supported by the NSF AI Institute for Foundations of Machine Learning (IFML).

References

[AVD25] Vahan Arsenyan, Elen Vardanyan, and Arnak Dalalyan. Assessing the quality of denoising diffusion models in Wasserstein distance: Noisy score and optimal bounds. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2506.09681.

[BDBDD24] Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on Learning Representations*, 2024. arXiv:2308.03686.

[BDD24] Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024. arXiv:2305.16860.

[CCDR26] Fan Chen, Sinho Chewi, Constantinos Daskalakis, and Alexander Rakhlin. High-accuracy sampling for diffusion models and log-concave distributions. In *International Conference on Machine Learning*, 2026. arXiv:2602.01338.

[CCL+23a] Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. In *Advances in Neural Information Processing Systems*, volume 36, pages 68552–68575, 2023.

[CCL+23b] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023. arXiv:2209.11215.

[CDD23] Sitan Chen, Giannis Daras, and Alex Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for DDIM-type samplers. In *International Conference on Machine Learning*, pages 4462–4484. PMLR, 2023.

[CdDPL+24] Sinho Chewi, Jaume de Dios Pont, Jerry Li, Chen Lu, and Shyam Narayanan. Query lower bounds for log-concave sampling. *Journal of the ACM*, 71(4):29:1–29:42, 2024. Preliminary version in FOCS 2023; arXiv:2304.02599.

[CDGS25] Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. KL convergence guarantees for score diffusion models under minimal data assumptions. *SIAM Journal on Mathematics of Data Science*, 7(1):86–109, 2025. arXiv:2308.12240.

[CGL+22] Sinho Chewi, Patrik Gerber, Chen Lu, Thibaut Le Gouic, and Philippe Rigollet. The query complexity of sampling from strongly log-concave distributions in one dimension. In *Conference on Learning Theory*, 2022. arXiv:2105.14163.

[CGLL23] Sinho Chewi, Patrik Gerber, Holden Lee, and Chen Lu. Fisher information lower bounds for sampling. In *Algorithmic Learning Theory*, 2023. arXiv:2210.02482.

[CLL23] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, second edition, 2006.

[DB22] Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. Expert Certification.

[DPS23] Alex Dytso, H. Vincent Poor, and Shlomo Shamai (Shitz). Conditional mean estimation in Gaussian noise: A meta derivative identity with applications. *IEEE Transactions on Information Theory*, 69(3):1883–1898, 2023.

[Efr11] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

[FSI+25] Kenji Fukumizu, Taiji Suzuki, Noboru Isobe, Kazusato Oko, and Masanori Koyama. Flow matching achieves almost minimax optimal convergence. In *The Thirteenth International Conference on Learning Representations*, 2025.

[GCC25] Shivam Gupta, Linda Cai, and Sitan Chen. Faster diffusion sampling with randomized midpoints: Sequential and parallel. In *International Conference on Learning Representations*, 2025. arXiv:2406.00924.

[GCS26] Khashayar Gatmiry, Sitan Chen, and Adil Salim. High-accuracy and dimension-free sampling with diffusions. In *International Conference on Machine Learning*, 2026. arXiv:2601.10708.

[GLR18] Rong Ge, Holden Lee, and Andrej Risteski. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering Langevin Monte Carlo. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[GNZ25] Xuefeng Gao, Hoang M. Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, 26(43):1–54, 2025. arXiv:2311.11003.

[GPPX24] Shivam Gupta, Aditya Parulekar, Eric Price, and Zhiyang Xun. Improved sample complexity bounds for diffusion model training. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 40976–41012. Curran Associates, Inc., 2024.

[GSO25] Marta Gentiloni Silveri and Antonio Ocello. Beyond log-concavity and score regularity: Improved convergence bounds for score-based generative models in W_2-distance. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 55631–55656, 2025. arXiv:2501.02298.

[GZ25] Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow ODEs of diffusion models in Wasserstein distances. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pages 1009–1017, 2025. arXiv:2401.17958.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[Hyv05] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

[HZ25] Yuchen He and Chihao Zhang. On the query complexity of sampling from non-log-concave distributions (extended abstract). In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 2786–2787. PMLR, 2025. Full version: arXiv:2502.06200.

[JL25] Yuchen Jiao and Gen Li. Instance-dependent convergence theory for diffusion models. *arXiv preprint arXiv:2410.13738v2*, 2025.

[JZL25] Yuchen Jiao, Yuchen Zhou, and Gen Li. Optimal convergence analysis of DDPM for general distributions. *arXiv preprint arXiv:2510.27562*, 2025.

[KAAL22] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

[LCBH+23] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. arXiv:2210.02747.

[LGL23] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. arXiv:2209.03003.

[LHE+24] Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, provably. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. arXiv:2403.03852.

[LJ25] Gen Li and Yuchen Jiao. Improved convergence rate for diffusion probabilistic models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[LLT22] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2206.06227.

[LLT23] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

[LRLZ22] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022. arXiv:2202.09778.

[LWCC24] Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. A sharp convergence theory for the probability flow ODEs of diffusion models. *arXiv preprint arXiv:2408.02320*, 2024.

[LY25] Gen Li and Yuling Yan. O(d/T) convergence theory for diffusion probabilistic models under minimal assumptions. *Journal of Machine Learning Research*, 26(292):1–55, 2025. arXiv:2409.18959.

[LZB+22] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, 2022.

[LZB+25] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22(4):730–751, 2025. arXiv:2211.01095.

[ND21] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[Rob56] Herbert Robbins. An empirical Bayes approach to statistics. In Jerzy Neyman, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 157–163, Berkeley, CA, 1956. University of California Press.

[SDCS23] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252, 2023. arXiv:2303.01469.

[SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, 2015.

[SE19] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

[SH22] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. arXiv:2202.00512.

[SME21] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[SSDK+21] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. arXiv:2011.13456.

[TY24] Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1648–1656. PMLR, 02–04 May 2024.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[Vin11] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[WCW24] Yuchen Wu, Yuxin Chen, and Yuting Wei. Stochastic Runge–Kutta methods: Provable acceleration of diffusion models. *arXiv preprint arXiv:2410.04760*, 2024.

[XGP25] Zhiyang Xun, Shivam Gupta, and Eric Price. Posterior sampling by combining diffusion models with annealed Langevin dynamics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[ZBR+23] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2302.04867.

[ZC23] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *International Conference on Learning Representations*, 2023. arXiv:2204.13902.

[ZHH+25] Matthew S. Zhang, Stephen Huan, Jerry Huang, Nicholas M. Boffi, Sitan Chen, and Sinho Chewi. Sublinear iterations can suffice even for DDPMs. *arXiv preprint arXiv:2511.04844*, 2025.

A Setup

Sections A–D prove Theorem 2.1, following the roadmap in Section 2.2. Extensions to non-uniform weights, subexponential errors, and Wasserstein distance follow in Sections E–G. Appendix I describes the synthetic illustration in Figure 3.

As in the main text, for a target law π and $\sigma > 0$, write

$$\pi_\sigma := \pi * \mathcal{N}(0, \sigma^2 I_d), \qquad s_{\pi,\sigma} := \nabla \log \pi_\sigma.$$

For an algorithm $\mathcal{A}$ interacting with $(\pi, \widehat{s})$, write $\widehat{X}^{\mathcal{A}}_{\pi,\widehat{s}}$ for its output.

It suffices to prove Theorem 2.1 for $p \geq 2$: for $0 < p < 2$, the $p = 2$ construction satisfies the required moment bound by Jensen's inequality. Fix the parameters $p, \rho, a, R, \gamma, d, \varepsilon_{\mathrm{err}}$ from that theorem, with $p \geq 2$, and write

$$H := \log(d/\varepsilon_{\mathrm{err}}) + \log(R/\gamma), \qquad B_d := \frac{d \log(R/\gamma)}{\sqrt{dH} + H}.$$

Throughout the analytic estimates, $c, C > 0$ denote universal constants that may change from line to line; subscripts indicate the permitted parameter dependence. Fix throughout

$$\delta_d := \frac{\rho^2}{8d^a}, \qquad w := A_{p,\rho,a}\big(\sqrt{dH} + H\big),$$

where $A_{p,\rho,a} > 0$ will be chosen sufficiently large.

B The target–oracle family

B.1 Product-circle mixtures and the common null reference

Let

$$b := \lfloor d/2 \rfloor, \qquad M := \left\lceil \frac{\pi R}{\gamma} \right\rceil.$$

Place M equally spaced points on the circle of radius R:

$$v_k := R\big(\cos(2\pi k/M), \sin(2\pi k/M)\big), \qquad 0 \leq k < M.$$

Let

$$V := \{v_0, \ldots, v_{M-1}\}^b \times \{0\}^{d-2b} \subset [-R, R]^d, \qquad \nu_{\mathrm{null}} := \mathrm{Unif}(V).$$

There are b planar blocks and, for odd d, one fixed coordinate. For an integer $n \geq 1$, draw an i.i.d. codebook

$$S = (Y_1, \ldots, Y_n), \qquad Y_i \overset{\mathrm{iid}}{\sim} \nu_{\mathrm{null}},$$

and let

$$\nu_S := \frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i}.$$

At total noise level $\tau > 0$, define the planted and null Gaussian mixtures

$$\pi_{S,\tau} := \nu_S * \mathcal{N}(0, \tau^2 I_d), \qquad \pi_{\mathrm{null},\tau} := \nu_{\mathrm{null}} * \mathcal{N}(0, \tau^2 I_d).$$

Their densities are denoted by $f_{S,\tau}$ and $f_{\mathrm{null},\tau}$, and their scores by $s_{S,\tau}$ and $s_{\mathrm{null},\tau}$. Here τ denotes total noise; Section B.3 relates it to the added smoothing σ.

The full product support has

$$\log|V| = b\log M \asymp d\log(R/\gamma),$$

which determines the range of log-codebook sizes used below.

The null mixture is common to every codebook. For $y \in V$ and $x \in \mathbb{R}^d$, define the log-likelihood ratio of one codeword against the null mixture by

$$\imath_\tau(y, x) := \log\frac{\varphi_{\tau^2}(x-y)}{f_{\mathrm{null},\tau}(x)},$$

where φ_{τ^2} is the density of $\mathcal{N}(0, \tau^2 I_d)$. If $Y \sim \nu_{\mathrm{null}}$ and $Z \sim \mathcal{N}(0, I_d)$ are independent, write

$$\imath_\tau := \imath_\tau(Y, Y+\tau Z), \qquad I_\tau := \mathbb{E}\,\imath_\tau = I(Y; Y+\tau Z).$$

Define the empirical log-likelihood ratio

$$r_{S,\tau}(x) := \log\frac{f_{S,\tau}(x)}{f_{\mathrm{null},\tau}(x)}.$$

Then

$$e^{r_{S,\tau}(x)} = \frac{1}{n}\sum_{i=1}^{n} e^{\imath_\tau(Y_i,x)}, \qquad \nabla r_{S,\tau} = s_{S,\tau} - s_{\mathrm{null},\tau}. \tag{1}$$

In particular, for every fixed x,

$$\mathop{\mathbb{E}}_{S} e^{r_{S,\tau}(x)} = 1. \tag{2}$$

B.2 Score bounds and information concentration

Lemma B.1 (Gaussian smoothing and score bounds)**.** *Let $Y \sim \nu$ be supported on $[-R, R]^d$, let $Z \sim \mathcal{N}(0, I_d)$ be independent of Y, and set $X := Y + \tau Z$. Write $f_{\nu,\tau}$ for the density of X and $s_{\nu,\tau} := \nabla \log f_{\nu,\tau}$ for its score. Then*

$$s_{\nu,\tau}(x) = \frac{\mathbb{E}[Y \mid X = x] - x}{\tau^2}$$

and

$$\nabla s_{\nu,\tau}(x) = -\frac{1}{\tau^2} I_d + \frac{1}{\tau^4}\operatorname{Cov}(Y \mid X = x).$$

Consequently, for any two such input laws ν_1, ν_2, $x \in \mathbb{R}^d$,

$$\|s_{\nu_1,\tau}(x) - s_{\nu_2,\tau}(x)\|_2 \le \frac{2R\sqrt{d}}{\tau^2},$$

and

$$\operatorname{Lip}(s_{\nu,\tau}) \le \frac{1}{\tau^2} + \frac{R^2 d}{\tau^4}.$$

Proof. The two displayed identities are the vector Tweedie and Hatsell–Nolte identities [DPS23]. Since every posterior mean has norm at most $R\sqrt{d}$, two posterior means differ by at most $2R\sqrt{d}$. Moreover,

$$\|\operatorname{Cov}(Y \mid X = x)\|_{\mathrm{op}} \le R^2 d,$$

which gives the score Lipschitz bound. □

Lemma B.2 (Noise-uniform concentration of information density)**.** *There is a universal constant $c > 0$ such that, for every $\tau > 0$ and every $u \geq 0$,*

$$\mathbb{P}\left[|\imath_\tau - I_\tau| \geq u\right] \leq 2\exp\left[-c\min\left\{\frac{u^2}{d}, u\right\}\right].$$

Proof. The channel factors across the b planar blocks, and the possible last coordinate cancels from the likelihood ratio. Thus $\imath_\tau$ is a sum of independent block contributions. By rotational symmetry, a block contribution has the distribution

$$\log M - \log \mathcal{Z}_\tau(Z), \qquad Z \sim \mathcal{N}(0, I_2),$$

where

$$\mathcal{Z}_\tau(z) := \sum_{k=0}^{M-1} \exp\left(-\frac{\|v_0 - v_k\|_2^2}{2\tau^2} - \frac{\langle z, v_0 - v_k\rangle}{\tau}\right).$$

The circle chord bounds give $\|v_k - v_0\|_2 \asymp \gamma \min\{k, M-k\}$. Hence the number of points within distance τ of v_0 satisfies

$$K_\tau := \#\{k : \|v_k - v_0\|_2 \leq \tau\} \asymp \min\{M, 1 + \tau/\gamma\},$$

and comparison with a Gaussian integral gives

$$\sum_k e^{-\|v_k - v_0\|_2^2/(4\tau^2)} \leq CK_\tau.$$

Keeping the K_τ nearby terms for the lower bound and completing the square for the upper bound yields

$$K_\tau e^{-1/2 - \|z\|_2} \leq \mathcal{Z}_\tau(z) \leq e^{\|z\|_2^2} \sum_k e^{-\|v_k - v_0\|_2^2/(4\tau^2)} \leq CK_\tau e^{\|z\|_2^2}.$$

Thus $|\log \mathcal{Z}_\tau(Z) - \log K_\tau| \leq C(1 + \|Z\|_2^2)$, so each centered block contribution has a universal subexponential norm. Bernstein's inequality [Ver18, Theorem 2.8.1] applied to the b blocks gives the stated bound. □

B.3 The masked oracle

When the planted score is already close to the null score, the oracle answers with the null score everywhere. Otherwise, it reveals the planted correction only on a high-likelihood region. Write

$$G_\tau := \frac{2R\sqrt{d}}{\tau^2},$$

the pointwise score-gap bound from Lemma B.1. Choose a 1-Lipschitz cutoff $\chi : \mathbb{R} \to [0, 1]$ with $\chi(u) = 0$ for $u \leq 0$ and $\chi(u) = 1$ for $u \geq 1$, and define the L^p mask

$$\psi_{S,\tau}(x) := \chi\left(r_{S,\tau}(x) + \log n - I_\tau + w/2 + 1\right).$$

Lemma B.3 (Mask properties)**.** *For sufficiently large $A_{p,\rho,a}$, every $\tau > 0$ and every $n \geq 1$, the following hold. For every fixed codebook $S \in V^n$,*

$$\pi_{S,\tau}(\psi_{S,\tau} < 1) \leq 2\exp\left[-c\min\left\{\frac{w^2}{d}, w\right\}\right] \leq \left(\frac{\varepsilon_{\mathrm{err}}\gamma}{2R\sqrt{d}}\right)^p,$$
$$\mathrm{Lip}(\psi_{S,\tau}) \leq G_\tau.$$

For every fixed $x \in \mathbb{R}^d$ and an i.i.d. codebook $S \sim \nu_{\mathrm{null}}^{\otimes n}$,

$$\mathbb{P}_S[\psi_{S,\tau}(x) > 0] \leq \exp(1 + \log n - I_\tau + w/2).$$

Proof. If $X = Y^\star + \tau Z \sim \pi_{S,\tau}$, then the self term in (1) shows that

$$\{\psi_{S,\tau}(X) < 1\} \subseteq \{\imath_\tau(Y^\star, X) < I_\tau - w/2\}.$$

By symmetry, the information-density tail is the same for every fixed codeword. Lemma B.2 gives the coverage bound, using $\log(2R\sqrt{d}/(\varepsilon_{\mathrm{err}}\gamma)) \leq CH$ for the last inequality. The Lipschitz bound follows from $\mathrm{Lip}(r_{S,\tau}) \leq G_\tau$. Finally, $\psi_{S,\tau}(x) > 0$ requires $r_{S,\tau}(x) > I_\tau - \log n - w/2 - 1$, so (2) and Markov's inequality give the blind-query bound. □

Let

$$\mathcal{M}_{p,\tau}(S) := \mathop{\mathbb{E}}_{X\sim\pi_{S,\tau}} \|\nabla r_{S,\tau}(X)\|_2^p.$$

The oracle is

$$\widehat{s}_{S,\tau}(x) := \begin{cases} s_{\mathrm{null},\tau}(x), & \mathcal{M}_{p,\tau}(S) \le (\varepsilon_{\mathrm{err}}/\tau)^p, \\ s_{\mathrm{null},\tau}(x) + \psi_{S,\tau}(x)\nabla r_{S,\tau}(x), & \mathcal{M}_{p,\tau}(S) > (\varepsilon_{\mathrm{err}}/\tau)^p. \end{cases}$$

Lemma B.4 (The oracle is valid)**.** *For every fixed codebook S and every $\tau \ge \gamma$,*

$$\mathop{\mathbb{E}}_{X\sim\pi_{S,\tau}} \|\widehat{s}_{S,\tau}(X) - s_{S,\tau}(X)\|_2^p \le \frac{\varepsilon_{\mathrm{err}}^p}{\tau^p},$$

and

$$\mathrm{Lip}(\widehat{s}_{S,\tau}) \le \frac{1}{\tau^2} + \frac{5R^2 d}{\tau^4}.$$

Proof. The null branch has the required error moment by definition. In the masked branch, the error is $-(1-\psi_{S,\tau})\nabla r_{S,\tau}$, so Lemma B.3 gives

$$\mathop{\mathbb{E}}_{\pi_{S,\tau}} \|\widehat{s}_{S,\tau} - s_{S,\tau}\|_2^p \le G_\tau^p \left(\frac{\varepsilon_{\mathrm{err}}\gamma}{2R\sqrt{d}}\right)^p = \frac{\varepsilon_{\mathrm{err}}^p \gamma^p}{\tau^{2p}} \le \frac{\varepsilon_{\mathrm{err}}^p}{\tau^p}.$$

For regularity, write the masked oracle as $(1-\psi_{S,\tau})s_{\mathrm{null},\tau} + \psi_{S,\tau}s_{S,\tau}$. Lemma B.1 and $\mathrm{Lip}(\psi_{S,\tau}) \le G_\tau$ give, by the product rule,

$$\mathrm{Lip}(\widehat{s}_{S,\tau}) \le \frac{1}{\tau^2} + \frac{R^2 d}{\tau^4} + G_\tau^2 = \frac{1}{\tau^2} + \frac{5R^2 d}{\tau^4}.$$

The null branch satisfies the same bound by Lemma B.1. □

Admissibility in the theorem's notation. For a fixed codebook S, set $\pi_S := \pi_{S,\gamma}$ and $\pi_{\mathrm{null}} := \pi_{\mathrm{null},\gamma}$. For $\sigma > 0$, put $\tau = \sqrt{\gamma^2 + \sigma^2}$ and define $\widehat{s}_\sigma^S := \widehat{s}_{S,\tau}$. Gaussian convolution gives

$$(\pi_S)_\sigma = \pi_{S,\tau}, \qquad s_{\pi_S,\sigma} = s_{S,\tau}, \qquad s_{\pi_{\mathrm{null}},\sigma} = s_{\mathrm{null},\tau}.$$

Since $\tau \ge \sigma$, Lemmas B.1 and B.4 give the accuracy and Lipschitz bounds in items (2)–(3) of Theorem 2.1; item (1) follows from $\pi_S = \nu_S * \mathcal{N}(0, \gamma^2 I_d)$. The map $\sigma \mapsto \tau$ is injective, so it also preserves the number of distinct query levels.

C The informative window is narrow

The informative window is centered at $\log n = I_\tau$. When $\log n$ is below this scale, the likelihood mask hides the signal from a fixed query. When $\log n$ is above it, averaging over the codebook makes the planted score close to the null score. We first prove the latter estimate in L^2.

Lemma C.1 (Dense codebooks are score-close)**.** *There are universal constants $c, C > 0$ such that, for every $\tau > 0$ and every integer $n \ge 1$ with $\Gamma := \log n - I_\tau > 0$,*

$$\mathop{\mathbb{E}}_{S} \mathcal{M}_{2,\tau}(S) \le C G_\tau^2 \exp\left[-c \min\left\{\frac{\Gamma^2}{d}, \Gamma\right\}\right].$$

Proof. Fix x, and let $m_{\mathrm{null},\tau}(x)$ be the null posterior mean. Write

$$W_i := e^{\iota_\tau(Y_i,x)}, \qquad u_i := \frac{Y_i - m_{\mathrm{null},\tau}(x)}{2R\sqrt{d}}, \qquad D := \frac{1}{n}\sum_i W_i, \quad \bar{u} := \frac{1}{n}\sum_i W_i u_i.$$

Then $\|u_i\|_2 \le 1$, $\mathbb{E}\, W_i = 1$, and $\mathbb{E}\, W_i u_i = 0$. Tweedie's formula gives $\nabla r_{S,\tau}(x) = G_\tau \bar{u}/D$. Since $f_{S,\tau}(x) = f_{\mathrm{null},\tau}(x) D$,

$$\mathop{\mathbb{E}}_{S} \mathcal{M}_{2,\tau}(S) = G_\tau^2 \int f_{\mathrm{null},\tau}(x) \mathop{\mathbb{E}}_{S} \frac{\|\bar{u}\|_2^2}{D}\, dx. \tag{3}$$

Truncate the weights at $T := e^{I_\tau + \Gamma/2}$. Let $D_\ell, \bar{u}_\ell$ be the same sums with only $W_i \le T$ retained, still normalized by n, and let $D_h := D - D_\ell$. Then

$$\frac{\|\bar{u}\|_2^2}{D} \le 4\|\bar{u}_\ell\|_2^2 + 2D_h + 2(1 - D_\ell)^2.$$

Indeed, if $D \ge 1/2$, use the squared triangle inequality and $\|\bar{u} - \bar{u}_\ell\|_2 \le D_h \le D$. If $D < 1/2$, the left side is at most $D < 1/2 \le 2(1 - D_\ell)^2$.

Put $q(x) := \mathbb{E}[W_i \mathbf{1}_{\{W_i > T\}}] \le 1$. Independence and $W_i^2 \mathbf{1}_{\{W_i \le T\}} \le T W_i$ imply

$$\mathbb{E}\,\|\bar{u}_\ell\|_2^2 \le T/n + q(x)^2, \qquad \mathbb{E}(1 - D_\ell)^2 \le T/n + q(x)^2, \qquad \mathbb{E}\,D_h = q(x).$$

Here the bias of each truncated average is at most $q(x)$, since $\mathbb{E}\,W_i u_i = 0$ and $\mathbb{E}\,W_i = 1$. Consequently, $\mathbb{E}_S[\|\bar{u}\|_2^2/D] \le C(T/n + q(x))$. Finally, $T/n = e^{-\Gamma/2}$, and a change of measure gives

$$\int f_{\mathrm{null},\tau}(x) q(x)\, dx = \mathbb{P}[\imath_\tau > I_\tau + \Gamma/2].$$

Substituting into (3) and applying Lemma B.2 proves the claim. □

Proposition C.2 (A narrow informative window)**.** *For a sufficiently large $A_{p,\rho,a}$, every $\tau \ge \gamma$ and every integer $n \ge 1$ satisfy the following for an i.i.d. codebook S of size n.*

(a) Sparse side. *If*

$$\log n \le I_\tau - w,$$

then, for every fixed $x \in \mathbb{R}^d$,

$$\mathbb{P}_S\left[\widehat{s}_{S,\tau}(x) \ne s_{\mathrm{null},\tau}(x)\right] \le \delta_d.$$

(b) Dense side. *If*

$$\log n \ge I_\tau + w,$$

then

$$\mathbb{P}_S\left[\widehat{s}_{S,\tau}(\,\cdot\,) \equiv s_{\mathrm{null},\tau}(\,\cdot\,)\right] \ge 1 - \delta_d.$$

Part (a) does not assert simultaneous agreement at all query points: its fixed-query bound suffices for the first-divergence coupling below. Part (b) instead puts the entire oracle in its null branch.

Proof. On the sparse side, a non-null answer requires $\psi_{S,\tau}(x) > 0$. By Lemma B.3,

$$\mathbb{P}_S[\widehat{s}_{S,\tau}(x) \ne s_{\mathrm{null},\tau}(x)] \le e^{1 + \log n - I_\tau + w/2} \le e^{1 - w/2} \le \delta_d.$$

On the dense side, $\mathcal{M}_{p,\tau}(S) \le G_\tau^{p-2} \mathcal{M}_{2,\tau}(S)$, since $p \ge 2$. Lemma C.1 and Markov's inequality give

$$\mathbb{P}_S[\mathcal{M}_{p,\tau}(S) > (\varepsilon_{\mathrm{err}}/\tau)^p] \le C \left(\frac{\tau G_\tau}{\varepsilon_{\mathrm{err}}}\right)^p \exp\left[-c \min\left\{\frac{\Gamma^2}{d}, \Gamma\right\}\right], \qquad \Gamma := \log n - I_\tau.$$

Since $\tau \ge \gamma$, the logarithm of the prefactor, including $1/\delta_d$, is at most $C_{p,\rho,a} H$. For $\Gamma \ge w$, this is absorbed by the exponent when $A_{p,\rho,a}$ is sufficiently large, giving failure probability at most δ_d. □

D From oracle indistinguishability to sampling lower bounds

At any fixed noise level, Proposition C.2 allows only a short interval of log-codebook sizes to affect the oracle. We therefore draw the size from a log-uniform prior: every short interval has small prior mass. Comparing the planted execution with the common null run handles adaptivity; a likelihood witness then gives the sampling lower bound.

D.1 Coupling to the common null execution

Set

$$m_d := \frac{b}{16}\log M, \qquad n_{\max} := \lfloor e^{m_d} \rfloor.$$

The factor $1/16$ leaves $\log n + m_d$ a linear margin below I_γ for $n \le n_{\max}$, using $I_\gamma \ge (b/5)\log M$, proved in Lemma D.2.

Draw

$$U \sim \mathrm{Unif}[0, m_d], \qquad N := \lfloor e^U \rfloor,$$

and, conditional on $N = n$, draw

$$S = (Y_1, \dots, Y_n), \qquad Y_i \overset{\text{iid}}{\sim} \nu_{\text{null}}.$$

For an algorithm $\mathcal{A}$, abbreviate its planted output by

$$\widehat{X}_S^{\mathcal{A}} := \widehat{X}^{\mathcal{A}}_{\pi_S, \widehat{s}^S}.$$

Let $\mathcal{D}$ be the induced law of $(\pi_S, \{\widehat{s}_\sigma^S\}_{\sigma>0})$. This law is independent of the algorithm.

Proposition D.1 (Level-sensitive null coupling)**.** *Let $\mathcal{A}$ be any randomized adaptive algorithm making at most Q queries and using at most ℓ distinct noise levels on every execution path. Couple its planted and null executions using the same internal randomness Ξ, independent of S and including its output randomness, and let* Div *be the event that the two executions differ. Then*

$$\mathbb{P}_{S,\Xi}(\mathsf{Div}) \le \frac{C\ell w}{m_d} + Q\delta_d.$$

Proof. Fix $\Xi = \xi$ and run $\mathcal{A}$ against $(\pi_{\text{null}}, \{s_{\pi_{\text{null}},\sigma}\}_{\sigma>0})$. Convert its query levels to $\tau = \sqrt{\gamma^2 + \sigma^2}$ as in Section B.3. This run contains at most Q queries and at most ℓ distinct levels. Since $0 \le U - \log N \le \log 2$, the event $|\log N - I_\tau| \le w$ confines U to an interval of length $O(w)$. Thus the probability that any null-run level is informative is at most $C\ell w/m_d$.

Outside this event, couple the planted and null executions using the same ξ. If they diverge, then immediately before their first differing answer they have the same transcript and ask one of the fixed null-run queries. That query is independent of S and lies outside the informative window, so Proposition C.2 bounds its disagreement probability by δ_d. A union bound over the at most Q possible first differing queries, followed by averaging over (Ξ, N), proves the bound. □

D.2 A target-only likelihood witness

Lemma D.2 (A clipped likelihood witness)**.** *For all sufficiently large d, depending only on ρ, and every codebook $S \in V^n$ with $1 \le n \le n_{\max}$, the function*

$$g_S(x) := \min\left\{1, \frac{(r_{S,\gamma}(x))_+}{m_d}\right\}$$

satisfies

$$\mathop{\mathbb{E}}_{\pi_S} g_S \ge 1 - \frac{\rho}{2}.$$

Moreover, for every fixed $x \in \mathbb{R}^d$ and $S \sim \nu_{\text{null}}^{\otimes n}$,

$$\mathop{\mathbb{E}}_{S} g_S(x) \le \frac{1}{m_d},$$

uniformly over $1 \le n \le n_{\max}$.

Proof. Nearest-neighbor decoding in one planar block has error probability at most

$$2\overline{\Phi}\left(\frac{R}{\gamma}\sin(\pi/M)\right) \le 2\overline{\Phi}(4/5) < 7/16,$$

where $\overline{\Phi}(t) := \mathbb{P}_{G\sim\mathcal{N}(0,1)}[G \geq t]$. Fano's inequality [CT06, Theorem 2.10.1 and the following corollary] therefore gives

$$I_\gamma \geq b\left(\frac{9}{16}\log M - \log 2\right) \geq \frac{b}{5}\log M,$$

where the last inequality uses $M \geq 7$.

If $X = Y^\star + \gamma Z \sim \pi_S$, the self term gives $r_{S,\gamma}(X) \geq \imath_\gamma(Y^\star, X) - \log n$. The choice of $n_{\max}$ leaves the margin

$$I_\gamma - \log n - m_d \geq \left(\frac{1}{5} - \frac{1}{8}\right) b \log M = \frac{3b}{40}\log M.$$

By symmetry and Lemma B.2, $\pi_S(r_{S,\gamma} < m_d) \leq 2e^{-cd} \leq \rho/2$ for large d, for every fixed S. This proves the target expectation bound.

For a fixed x, use $t_+ \leq e^t$ and (2):

$$\mathop{\mathbb{E}}_S g_S(x) \leq \frac{\mathbb{E}_S(r_{S,\gamma}(x))_+}{m_d} \leq \frac{\mathbb{E}_S\, e^{r_{S,\gamma}(x)}}{m_d} = \frac{1}{m_d}.$$

□

Proof of Theorem 2.1. Section B.3 verifies items (1)–(3) for every realization of the hard law constructed in Section D.1. Since $m_d \asymp d\log(R/\gamma)$, we have $B_d w/m_d = O_{p,\rho,a}(1)$. Choose the theorem's resource constant $c = c(p,\rho,a) \leq 1$ sufficiently small. Then any algorithm $\mathcal{A}$ allowed by item (4) has $\mathbb{P}_{S,\Xi}(\mathsf{Div}) \leq \rho^2/4$ by Proposition D.1. Let $\widehat{X}_0^{\mathcal{A}}$ be the null output in that coupling. It is independent of S, and equals $\widehat{X}_S^{\mathcal{A}}$ outside Div. Lemma D.2 and Fubini therefore give, for d sufficiently large,

$$\begin{aligned}\mathop{\mathbb{E}}_S\Big[\mathbb{E}\big[g_S(\widehat{X}_S^{\mathcal{A}}) \mid S\big]\Big] &\leq \mathbb{P}_{S,\Xi}(\mathsf{Div}) + \mathop{\mathbb{E}}_{S,\Xi} g_S(\widehat{X}_0^{\mathcal{A}}) \\ &\leq \frac{\rho^2}{4} + \frac{1}{m_d} \leq \frac{\rho^2}{2}.\end{aligned}$$

Thus, by Markov's inequality, with probability at least $1-\rho$ over S,

$$\mathbb{E}\big[g_S(\widehat{X}_S^{\mathcal{A}}) \mid S\big] \leq \rho/2.$$

For every such S, the variational characterization of total variation yields

$$\mathrm{d_{TV}}(\widehat{X}_S^{\mathcal{A}}, \pi_S) \geq \mathop{\mathbb{E}}_{\pi_S} g_S - \mathbb{E}\big[g_S(\widehat{X}_S^{\mathcal{A}}) \mid S\big] \geq 1-\rho.$$

This proves item (4). The "in particular" claim follows because a total budget cB_d gives both $Q \leq cB_d \leq d^a$ and $\ell \leq Q \leq cB_d$. All dimension thresholds used above depend only on $p, \rho, a, R/\gamma$. □

Derivation of Theorem 1.3. Take $a = 1$, $\rho = 0.01$, fixed constants $R > 2\gamma > 0$, and $\varepsilon_{\mathrm{err}} = d^{-O(1)}$ in Theorem 2.1. Then $B_d = \Omega(\sqrt{d/\log d})$, and the "in particular" clause gives Theorem 1.3.

Testing consequence. Give equal prior probability to the null transcript and to the transcript from a random planted instance. Proposition D.1 bounds the optimal testing advantage by $\frac{1}{2}[C\ell w/m_d + Q\delta_d]$. Thus advantage at least α, together with $Q\delta_d \leq \alpha$, requires

$$\ell = \Omega(\alpha m_d/w) = \Omega_{p,\rho,a}(\alpha B_d)$$

distinct levels.

E An extension to non-uniform weights

Proposition E.1 (Robustness to polynomially bounded reweighting)**.** *Fix a constant $C_{\mathrm{wt}} \geq 0$. For each $1 \leq n \leq n_{\max}$, fix a deterministic probability vector $\eta^{(n)} = (\eta_1^{(n)}, \dots, \eta_n^{(n)})$ satisfying*

$$n\|\eta^{(n)}\|_\infty \leq d^{C_{\mathrm{wt}}}.$$

Replace ν_S in the hard construction by $\nu_S^\eta := \sum_{i=1}^n \eta_i^{(n)} \delta_{Y_i}$, and let $\mathcal{D}^\eta$ be the induced weighted hard law. Then $\mathcal{D}^\eta$ witnesses Theorem 2.1 with the same B_d; the constants c and d_0 may additionally depend on C_{wt}. The result also holds for a random weight family drawn independently of U and all codeword locations and satisfying the displayed bound almost surely.

Proof. As before, it suffices to consider $p \geq 2$. Fix n, abbreviate $\eta := \eta^{(n)}$, and put $\eta_{\max} := \max_i \eta_i$. Use the weighted mixture in the likelihood ratio and oracle, enlarging the constant in w to depend also on C_{wt}. For every fixed x,

$$e^{r_{S,\tau}^\eta(x)} = \sum_i \eta_i e^{\imath_\tau(Y_i,x)}, \qquad \mathop{\mathbb{E}}_S e^{r_{S,\tau}^\eta(x)} = 1.$$

The sparse-side bound follows by Markov's inequality.

For mask coverage, generate a target sample as $X = Y_J + \tau Z$, where $J \sim \eta$. The self term gives $r_{S,\tau}^\eta(X) \geq \imath_\tau(Y_J, X) + \log \eta_J$, and $\mathbb{P}[\log(n\eta_J) < -t] \leq e^{-t}$: the small weights have total mass at most $n(e^{-t}/n)$. Combining this with information concentration, taking both deviations to be $h/2$, gives, for every fixed S and $h \geq 0$,

$$\pi_{S,\tau}^\eta\big(r_{S,\tau}^\eta + \log n < I_\tau - h\big) \leq Ce^{-c\min\{h^2/d,h\}}.$$

This bound gives mask coverage and the target witness estimate $\mathbb{E}_{\pi_S^\eta} g_S^\eta \geq 1 - \rho/2$, with the same m_d and $g_S^\eta := \min\{1, (r_{S,\gamma}^\eta)_+/m_d\}$. The mean-one likelihood identity gives $\mathbb{E}_S g_S^\eta(x) \leq 1/m_d$.

On the dense side, the variance bound T/n in the proof of Lemma C.1 becomes

$$T\sum_i \eta_i^2 \leq T\eta_{\max}.$$

The same truncation argument therefore applies with information gap

$$\log(1/\eta_{\max}) - I_\tau \geq \log n - I_\tau - C_{\mathrm{wt}} \log d.$$

Since $H \geq \log d$, enlarging the constant in w absorbs this loss and gives the dense-side bound with failure probability δ_d. Proposition D.1 and the witness argument now prove the theorem with a smaller resource constant c. All bounds are uniform over the allowed weight families, so conditioning on an independent random weight family and then averaging proves the last assertion. □

F Subexponential score errors

Only the oracle construction and the resulting window scale change from the L^p proof.

Theorem F.1 (Subexponential-error extension)**.** *Theorem 2.1 remains valid after removing the parameter p and making the following substitutions. Denote the oracle by $\widehat{s}^{\mathrm{se}}$, and replace item* (2) *by*

$$\mathbb{P}_{X\sim\pi_\sigma}\left[\|\widehat{s}_\sigma^{\mathrm{se}}(X) - s_{\pi,\sigma}(X)\|_2 \geq z\right] \leq 2\exp\left(-\frac{z\sigma}{\varepsilon_{\mathrm{err}}}\right) \qquad (\sigma > 0,\ z \geq 0).$$

In the resource bound, replace H and B_d by the quantities

$$H_{\mathrm{se}} := \frac{R\sqrt{d}}{\gamma\varepsilon_{\mathrm{err}}}, \qquad B_d^{\mathrm{se}} := \frac{d\log(R/\gamma)}{\sqrt{dH_{\mathrm{se}}} + H_{\mathrm{se}}}.$$

The constants may be chosen as $c = c(\rho, a) > 0$ and $d_0 = d_0(\rho, a, R/\gamma)$. All other conclusions of the theorem are unchanged.

Proof. Use the target family and size prior from Section D.1. In the mask and oracle construction, replace w by

$$w_{\mathrm{se}} := A_{\rho,a}^{\mathrm{se}}\big(\sqrt{dH_{\mathrm{se}}} + H_{\mathrm{se}}\big),$$

and use the globally null branch exactly when

$$\mathop{\mathbb{E}}_{\pi_{S,\tau}} e^{\tau\|\nabla r_{S,\tau}\|_2/\varepsilon_{\mathrm{err}}} \leq 2.$$

Write $\psi^{\mathrm{se}}_{S,\tau}$ and $\widehat{s}^{\mathrm{se}}_{S,\tau}$ for this mask and oracle. The Lipschitz bound follows as in Lemma B.4.

Put

$$u_\tau := \frac{\tau G_\tau}{\varepsilon_{\mathrm{err}}} = \frac{2R\sqrt{d}}{\tau\varepsilon_{\mathrm{err}}}.$$

At total noise $\tau \geq \gamma$, the error's exponential moment is at most 2 in the null branch by definition. For the masked branch, the coverage argument in Lemma B.3 gives $\pi_{S,\tau}(\psi^{\mathrm{se}}_{S,\tau} < 1) \leq e^{-u_\gamma}$ for sufficiently large $A^{\mathrm{se}}_{\rho,a}$, since $u_\gamma = 2H_{\mathrm{se}}$. The error is at most G_τ and vanishes wherever the mask is one, so

$$\mathop{\mathbb{E}}_{\pi_{S,\tau}} e^{\tau\|\widehat{s}^{\mathrm{se}}_{S,\tau} - s_{S,\tau}\|_2/\varepsilon_{\mathrm{err}}} \leq 1 + e^{u_\tau}\pi_{S,\tau}(\psi^{\mathrm{se}}_{S,\tau} < 1) \leq 2,$$

since $u_\tau \leq u_\gamma$. Exponential Markov gives the required tail in either branch.

The sparse-window proof also applies with w_{se}. For the dense side, $e^v - 1 \leq ve^{u_\tau}$ for $0 \leq v \leq u_\tau$, Markov's inequality, and Cauchy–Schwarz give

$$\begin{aligned}
\mathbb{P}_S\left[\mathop{\mathbb{E}}_{\pi_{S,\tau}} e^{\tau\|\nabla r_{S,\tau}\|_2/\varepsilon_{\mathrm{err}}} > 2\right] &\leq \mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\pi_{S,\tau}}\left[e^{\tau\|\nabla r_{S,\tau}\|_2/\varepsilon_{\mathrm{err}}} - 1\right] \\
&\leq \frac{\tau e^{u_\tau}}{\varepsilon_{\mathrm{err}}}\sqrt{\mathop{\mathbb{E}}_{S}\mathcal{M}_{2,\tau}(S)} \\
&\leq C\exp\left(CH_{\mathrm{se}} - c\min\{\Gamma^2/d, \Gamma\}\right),
\end{aligned}$$

where $\Gamma := \log n - I_\tau > 0$; the last line uses Lemma C.1 and $u_\tau \leq 2H_{\mathrm{se}}$. Since $\log(1/\delta_d) = O_{\rho,a}(H_{\mathrm{se}})$, $\Gamma \geq w_{\mathrm{se}}$ makes this probability at most δ_d after enlarging $A^{\mathrm{se}}_{\rho,a}$.

Proposition D.1 and Lemma D.2 give the level bound, since $m_d/w_{\mathrm{se}} \asymp_{\rho,a} B^{\mathrm{se}}_d$. Translate the oracle to σ as in Section B.3; the tail bound follows from $\sqrt{\gamma^2 + \sigma^2} \geq \sigma$. Since $B^{\mathrm{se}}_d \leq d \leq d^a$, the total-query conclusion follows as well. □

For fixed R/γ and $\varepsilon_{\mathrm{err}}$, we have $H_{\mathrm{se}} \asymp \sqrt{d}$ and $B^{\mathrm{se}}_d \asymp d^{1/4}$. Applying the theorem with $a = 1$, $\rho = 0.01$, fixed $R > 2\gamma > 0$, and accuracy $\min\{\varepsilon_{\mathrm{err}}, 1\}$ proves Theorem 1.6 for every constant $\varepsilon_{\mathrm{err}} > 0$.

G A Wasserstein extension

Rescaling the clipped likelihood witness in Lemma D.2 gives a 1-Lipschitz test function.

Corollary G.1 (Wasserstein lower bound)**.** *In the setting of Theorem 2.1, for every $q \in [1, \infty)$, the same hard law $\mathcal{D}$ constructed in Section D.1 satisfies*

$$\mathbb{P}_{(\pi,\widehat{s})\sim\mathcal{D}}\left[W_q(\widehat{X}^{\mathcal{A}}_{\pi,\widehat{s}}, \pi) \geq c_W\frac{\gamma^2\log(R/\gamma)}{R}\sqrt{d}\right] \geq 1 - \rho$$

for every algorithm obeying the query budgets in item (4)*, where $c_W > 0$ is universal. The total-query conclusion also holds.*

Proof. Use the same resource constant and high-probability witness bound as in the proof of Theorem 2.1. Since $\mathrm{Lip}(r_{S,\gamma}) \leq G_\gamma$, $\phi_S := (m_d/G_\gamma)g_S$ is bounded and 1-Lipschitz. For each instance in that event, the coupling definition of W_1 gives

$$W_1(\widehat{X}^{\mathcal{A}}_S, \pi_S) \geq \mathop{\mathbb{E}}_{\pi_S}\phi_S - \mathbb{E}[\phi_S(\widehat{X}^{\mathcal{A}}_S) \mid S] \geq (1-\rho)\frac{m_d}{G_\gamma}.$$

Since $b \geq d/3$, $\log M \geq \log(R/\gamma)$, and $\rho < 1/4$,

$$(1-\rho)\frac{m_d}{G_\gamma} = (1-\rho)\frac{b\gamma^2\log M}{32R\sqrt{d}} \geq \frac{\gamma^2\log(R/\gamma)}{128R}\sqrt{d}.$$

Finally, $W_q \geq W_1$ for $q \geq 1$, with the extended-value interpretation when moments are infinite. □

Derivation of Theorem 1.5. Take $a = 1$, $\rho = 0.01$, fixed constants $R > 2\gamma > 0$, and $\varepsilon_{\mathrm{err}} = d^{-O(1)}$ in Corollary G.1. Then $B_d = \Omega(\sqrt{d/\log d})$, and the total-query conclusion gives Theorem 1.5 with $c = c_W\gamma^2\log(R/\gamma)/R$.

H Applicability to Practical Algorithms

Our oracle model is intended to capture the inference-time information available to a broad class of practical diffusion and diffusion-adjacent samplers: repeated evaluations of a learned network at user-chosen states and noise levels, without any additional access to the target distribution π.

Common network parameterizations are score-equivalent. In standard diffusion setups, a network may be parameterized as a denoiser/x_0-predictor, a noise predictor, or directly as a score/drift model. All of these are information-equivalent to a smoothed-score oracle, up to known, noise-dependent affine transformations. Concretely, under additive Gaussian corruption, the denoiser (posterior mean) $D_\sigma(x)$ and the noise predictor $\varepsilon_\sigma(x)$ satisfy the standard identities

$$s_{\pi,\sigma}(x) \;=\; \frac{D_\sigma(x) - x}{\sigma^2}, \qquad \varepsilon_\sigma(x) \;=\; -\sigma\, s_{\pi,\sigma}(x),$$

and hence any learned approximation of D_σ or ε_σ can be converted into a score estimate $\widehat{s}_\sigma$ by the same formulas. These conversions are routinely used in diffusion models such as DDPM/DDIM and score-SDE formulations (up to conventional rescalings that depend only on the noise/time parameterization) [HJA20, SSDK+21, KAAL22, SME21]. Therefore, up to constant factors, each neural network evaluation can be counted as one oracle query. (Methods that use higher-order updates or corrector steps may spend multiple network evaluations per macro-step, but the relevant complexity measure for our results is the *total* number of evaluations.)

SDE/ODE-based diffusion samplers. This scope includes both classic reverse-time SDE discretizations and ODE-based solvers. Concretely, it covers DDPM and its variants, predictor–corrector samplers, the probability flow ODE, and generic integrators such as DPM-Solver, PNDM, DEIS, UniPC, and EDM-style solvers [SSDK+21, LZB+22, LZB+25, LRLZ22, ZC23, ZBR+23, KAAL22]. These methods differ in discretization order, step-size selection, and whether the path is formulated as an SDE or an ODE, but they all obtain information about π through the same primitive: evaluating a learned score (or an equivalent drift derived from it) at adaptively chosen states and noise levels.

Flow matching and rectified flow in the linear Gaussian interpolation setting. Our oracle model also captures the standard rectified-flow / flow-matching setup with a Gaussian prior and linear interpolation. Specifically, let $X_1 \sim \pi$ and $X_0 \sim \mathcal{N}(0, I_d)$ be independent and define

$$X_t = tX_1 + (1-t)X_0, \qquad t \in (0,1).$$

Rectified flow defines the velocity field

$$v_t(x) := \mathbb{E}[X_1 - X_0 \mid X_t = x].$$

A direct calculation using Tweedie's formula [Rob56, Efr11] shows that v_t is an explicit affine transform of the Gaussian-smoothed score:

$$v_t(x) = \frac{x}{t} + \frac{1-t}{t^2}\, s_{\pi,\sigma_t}(x/t), \qquad \sigma_t = \frac{1-t}{t}.$$

Equivalently, for each fixed t, querying $v_t(x)$ is invertibly equivalent to querying s_{π,σ_t} at the rescaled input x/t. Therefore, any sampler whose inference-time access to π is only through evaluations of learned approximations of such velocity fields inherits the same query lower bounds [LCBH+23, LGL23, BDD24, FSI+25].

What is not covered. Approaches such as progressive distillation [SH22] or consistency models [SDCS23] may fall outside this model: their training procedures can encode additional global information about π into a new mapping, so that each inference-time evaluation is no longer a score-equivalent query. For hybrid systems that combine a partially distilled backbone with residual score-correction steps, applicability depends on whether the correction steps are the sole source of information about π at inference time; if so, the lower bound applies to the number of such correction queries.

I Synthetic experiments

The synthetic experiment in Figure 3 visualizes the location and width of the informative window on the hypercube warm-up family from Section 2.1. It is not a direct numerical evaluation of the theorem quantity; rather, it isolates, in a simplified model, the geometric mechanism behind the lower bound.

Setup. We work on the hypercube $\{-1,+1\}^d$ and fix the canonical query point $x = \mathbf{1}$. The experiment uses a Poissonized shell-count model: the occupancies $N_k := |\{y \in S : \mathrm{ham}(y, \mathbf{1}) = k\}|$ are sampled as independent Poisson random variables with means

$$\mu_k \;=\; \binom{d}{k} e^{\kappa d - d\log 2},$$

so that the expected codebook size is $e^{\kappa d}$. This is a standard approximation to the combinatorial random-coding model of Section 2.1 that becomes exact in the limit $d \to \infty$ at fixed κ; for the sparse regime ($\kappa = 0.2$) and dimensions used here, the two models behave indistinguishably.

Let $q_\kappa \in (0, 1/2)$ be the unique lower-branch solution of $h(q_\kappa) = \log 2 - \kappa$, where $h(q) = -q\log q - (1-q)\log(1-q)$ is the binary entropy. For $\kappa = 0.2$ one obtains $q_\kappa \approx 0.1948$.

Shell-resolved conditional proxy. The ideal pointwise signal at $x = \mathbf{1}$ is $\mathcal{I}_S(\tau; \mathbf{1})$, defined in the main text. Direct evaluation would require an explicit codebook of expected size $e^{\kappa d}$, which is infeasible at the dimensions considered here (d up to 32768). We therefore replace the realized signal by a shell-resolved proxy: we condition on the shell occupancies $\{N_k\}$ and analytically integrate out the randomness from vertex placement within each shell. This removes within-shell sampling noise and isolates the dependence on the shell occupancy profile.

For a smoothing level $\tau > 0$, write $u = e^{-2/\tau^2}$. A vertex y at Hamming distance $k = \mathrm{ham}(y, \mathbf{1})$ from $\mathbf{1}$ contributes a weight $\exp(-\|\mathbf{1} - y\|^2/(2\tau^2)) = u^k$, so in the proxy all quantities depend only on $\{N_k\}$. Define the partition function and density ratio

$$Z(\tau) \;=\; \sum_{k=0}^{d} N_k\, u^k, \qquad R_\tau \;=\; \frac{2^d}{|S|}\,\frac{Z(\tau)}{(1+u)^d}, \qquad |S| = \sum_{k=0}^{d} N_k.$$

The posterior mean per coordinate under the planted codebook, averaged over coordinates, is

$$\bar{\alpha}(\tau) \;=\; \frac{\sum_{k=0}^{d} N_k\, u^k\, (1 - 2k/d)}{Z(\tau)},$$

while the null posterior mean per coordinate is $m_0(\tau) = (1-u)/(1+u)$.

The formula for the proxy is derived as follows. Under a model in which, conditional on the shell counts, the N_k vertices in shell k form a uniformly random subset of that shell, the conditional expectation of $(1/d)\|m(\mathbf{1}) - m^{\mathrm{null}}(\mathbf{1})\|_2^2$ given $\{N_k\}$ decomposes into a *bias* term and a *variance* term:

$$G_\tau^{\mathrm{cond}} \;=\; \underbrace{(\bar{\alpha}(\tau) - m_0(\tau))^2}_{\text{shell-occupancy bias}} \;+\; \underbrace{\sum_{k=1}^{d-1} \frac{N_k\, u^{2k}}{Z(\tau)^2}\; v_k\, f_k}_{\text{within-shell variance}},$$

where

$$v_k = \frac{4k(d-k)}{d^2}, \qquad f_k = \frac{\binom{d}{k} - N_k}{\binom{d}{k} - 1}.$$

Here v_k is the per-coordinate variance of a uniformly random vertex in shell k, and f_k is the finite-population (hypergeometric) correction factor. The first term captures how the occupancy profile $\{N_k\}$ shifts the overall posterior mean away from the null; the second captures the residual fluctuation from the random placement of vertices within each shell. The boundary shells $k \in \{0, d\}$ contribute zero to the variance sum since each contains a single vertex.

The proxy for $\mathcal{I}_S(\tau; \mathbf{1})$ is then

$$\widetilde{\mathcal{I}}(\tau) \;=\; \tau^{-4}\, R_\tau\, G_\tau^{\mathrm{cond}},$$

where the factor τ^{-4} converts the posterior-mean difference to a score difference via Tweedie's formula. In the experiment, we evaluate this formula at shell counts drawn from the Poissonized model. All reported curves are *medians* of $\widetilde{\mathcal{I}}(\tau)$ over independently sampled shell-occupancy vectors. We use the median rather than the mean because the upper tails from rare favorable shell configurations make the median a more stable summary of typical behavior.

Heuristic prediction. The null shell weights satisfy

$$\frac{\binom{d}{k}\,u^k}{(1+u)^d} \;=\; \Pr\big[\mathrm{Bin}(d,q_\tau)=k\big], \qquad q_\tau=\frac{u}{1+u}=\frac{1}{1+e^{2/\tau^2}}.$$

Hence the shells contributing at smoothing level τ lie in a band $k=q_\tau d+O(\sqrt{d})$.

Recall that q_κ is the critical shell fraction at which the expected occupancy transitions from exponentially sparse to exponentially dense. By Stirling's formula,

$$\log\mu_{qd} \;=\; d\big(h(q)-\log 2+\kappa\big)-\tfrac{1}{2}\log\big(2\pi d\,q(1-q)\big)+O(1),$$

and since $h'(q_\kappa)=\log\frac{1-q_\kappa}{q_\kappa}>0$, for $q=q_\kappa+\delta/\sqrt{d}$ one has

$$\log\mu_{qd}-\log\mu_{q_\kappa d} \;=\; h'(q_\kappa)\,\delta\sqrt{d}+O(1+\delta^2).$$

Thus across a shell band of width $O(d^{-1/2})$ in q, the expected occupancies change by $e^{\Theta(\sqrt{d})}$: shells slightly inside q_κ are on the sparse side, while shells slightly outside q_κ are already on the dense side. The informative regime therefore occurs when the contributing binomial band is centered within $O(d^{-1/2})$ of q_κ, i.e. $q_\tau-q_\kappa=O(d^{-1/2})$.

Passing to $t=\log\tau$, the map $t\mapsto q_\tau$ is smooth with

$$\frac{dq_\tau}{d\log\tau} \;=\; 2\,q_\tau(1-q_\tau)\,\log\frac{1-q_\tau}{q_\tau},$$

which is nonzero at the matching scale τ_* defined by $q_{\tau_*}=q_\kappa$. Therefore the informative window on the $\log\tau$ axis also has width $\Theta(d^{-1/2})$. The figure fixes κ and varies τ, whereas Proposition C.2 fixes τ and varies $\log n$.

Numerical procedure. We evaluate the median curve on a uniform grid of 641 points in $\log\tau$ over the interval $[\log\tau_*-0.5,\ \log\tau_*+0.5]$, for

$$d\;\in\;\{512,\ 1024,\ 2048,\ 4096,\ 8192,\ 16384,\ 32768\}.$$

Shell occupancies are sampled as follows:

- $\mu_k<50$: exact Poisson draw;
- $50\le\mu_k<10^8$: Gaussian approximation $N_k=\max\{0,\ \mathrm{round}(\mu_k+\sqrt{\mu_k}\cdot Z)\}$, where $Z\sim\mathcal{N}(0,1)$;
- $\mu_k\ge 10^8$: deterministic replacement $N_k=\mu_k$.

All numerically delicate quantities—shell masses, weighted sums, and the density ratio—are accumulated in log-space. The right panel of Figure 3 reports the full width at half maximum (FWHM) of the raw median curve in the variable $\log\tau$, with the two half-height crossings obtained by linear interpolation.

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: The abstract and introduction state the $\widetilde{\Omega}(\sqrt{d})$ score-query lower bound under Assumptions 1 and 2; the formal statements appear as Theorem 1.3 and Theorem 1.6 and the scope of the model is described in Appendix H.

   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: The remaining gap to the best known $\widetilde{O}(d)$ upper bound, the dependence on the smoothed-score oracle, and the adversarial freedom of our hard instance are discussed throughout Appendix H and the Conclusion's future-work directions.

   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: All assumptions are stated in the main text alongside the theorems they support; full proofs are provided in the appendix and cross-referenced.

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 synthetic illustration (Figure 3) is fully specified in Appendix I, including the Poissonized shell-count sampling rules, the closed-form proxy formula, the $\log \tau$ grid, the dimensions $d \in \{512, \ldots, 32768\}$, and the FWHM computation.

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: [No]

   Justification: The paper is primarily theoretical and we do not release code; the synthetic illustration is reproducible from the analytical procedure given in Appendix I, which specifies the sampling distribution, the closed-form proxy, the $\log \tau$ grid, and the FWHM extraction.

   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: The synthetic illustration has no training; all numerical settings (Poisson sampling rules, log-space accumulation, 641-point $\log \tau$ grid, dimensions, FWHM via linear interpolation) are specified in Appendix I.

   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: We report median curves over independently sampled shell-occupancy vectors as a stable summary (the upper tail from rare favorable configurations makes the mean unstable); the predicted $1/\sqrt{d}$ scaling is supported by the consistent FWHM trend across seven dimensions $d \in \{512, \ldots, 32768\}$ rather than by per-curve error bars.

   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: The synthetic illustration is an analytical evaluation of a closed-form proxy on a 641-point $\log \tau$ grid for 7 dimensions and runs in minutes on a single CPU with no GPU or specialized hardware.

   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 work establishes information-theoretic lower bounds and contains only synthetic numerical illustrations; it conforms 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: The paper proves information-theoretic lower bounds for a class of sampling algorithms and proposes no model, dataset, or deployable method; we do not foresee a direct societal impact.

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: No models, datasets, or scraped resources are released.

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: [N/A]

Justification: The paper does not use external datasets, pretrained models, or third-party code; the synthetic illustration relies only on standard scientific Python libraries.

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: No new datasets, code, or models are released with the paper.

    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 or human-subjects research.

    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 human-subjects research and therefore no IRB approval is required.

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: LLMs were not used as a component of the methodology; any usage was limited to writing/editing assistance, for which declaration is not required.

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.