BEYOND QUERY COUNT: A RADIAL BUDGET FOR SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient diffusion sampling requires reducing the number of denoising model evaluations used to generate a sample. When a sampler can request estimates at several neighboring noise levels together, counting requests alone leaves their total evaluation cost unresolved. We study this question in a discrete model where noise flips a prescribed number of binary coordinates and each query returns estimates for a contiguous band of noise levels. We introduce a radial information budget that accounts for the overlap between neighboring levels. The key geometric bound charges a band for its width plus one logarithmic term, avoiding a separate logarithmic charge for every returned estimate. We carry this bound through complete adaptive response histories. For each declared evaluation cap, we construct random targets and accurate denoising responses for which approximate sampling requires a worst case radial budget linear in dimension. At fixed denoising accuracy and moment order, the resulting evaluation lower bound is proportional to dimension divided by its logarithm, even with adaptive query points, band widths, and stopping times. In this model, grouping neighboring estimates into fewer requests does not remove the lower bound on their total number.

1 INTRODUCTION

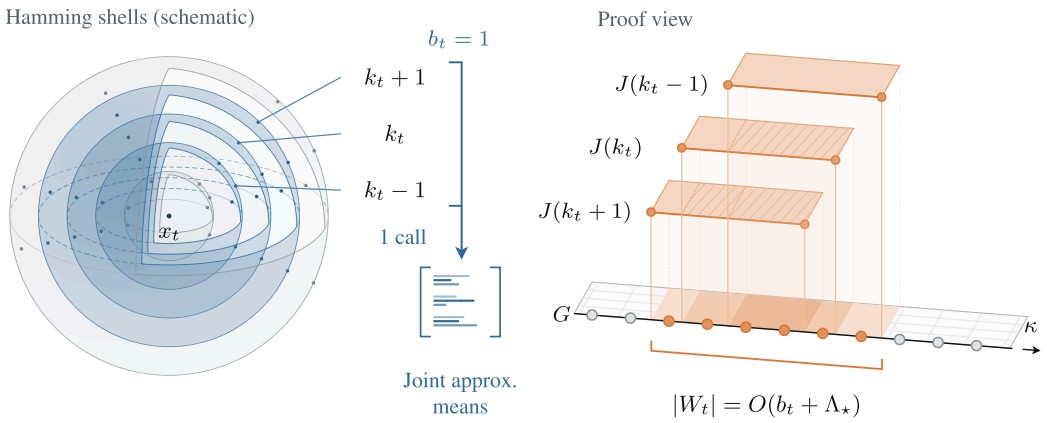
Diffusion models generate samples through repeated evaluations of a denoising model at different noise levels. These evaluations make sampling expensive: each step requires another model computation before generation can proceed. Designing samplers that retain sample quality with fewer evaluations has therefore become a central goal. Methods such as DPM-Solver (Lu et al., 2022) and improved sampling schedules (Karras et al., 2022) reduce this cost by choosing how to move between noise levels more efficiently.

How far can the number of evaluations be reduced? A lower bound addresses this question by identifying target distributions that require many evaluations, regardless of the sampling algorithm. Xun & Price (2026) established such bounds when each query returns an approximate denoising estimate at one selected noise level. Their Gaussian result shows that accurate estimates can still require many queries. It leaves a different resource question when a sampler can request estimates at several neighboring levels together: how many individual estimates are necessary, even if they arrive in fewer requests?

The difficulty is that neighboring estimates can reveal overlapping information about the target. Charging a separate information contribution for every estimate can count that overlap repeatedly, whereas counting only requests ignores how many estimates each request contains. A lower bound for this setting must control the contribution of the whole returned list, including when later requests depend on earlier answers.

We study this question in a discrete model that separates the number of noise levels requested from the accuracy of their estimates. The target is a uniform distribution over a random collection of binary vectors. A noise level specifies how many coordinates are flipped, so each level corresponds to one exact distance from the uncorrupted vector. A *band query* requests approximate denoising estimates at a contiguous range of these levels, all at the same query point. This fixed energy model makes the relation between neighboring levels explicit. Section 2 defines its accuracy criterion and distinguishes it from Gaussian denoising.

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



(a) One band query (b) Shared rate coverage

Figure 1: One band query and its overlapping rate coverage. (a) An interior $b_t = 1$ query selects three shells and returns one joint list of approximate posterior means. The shells and discrete states are schematic. (b) The analytical windows $J(k)$ for these noise levels overlap on a shared candidate grid G. Orange points form $W_t = G \cap \bigcup_{k \in K_{b_t}(k_t)} J(k)$, with $|W_t| = O(b_t + \Lambda_\star)$. These windows are proof objects, distinct from the returned means. Height and depth serve display roles. Hatching marks coverage from above, while darker floor regions count overlapping windows. Dashed arcs and traces show occluded or projected geometry. Outside the windows, the returned estimates rarely differ from the uniform background response. Appendix A details the graphical conventions.

Our main contribution is a *radial information budget* for adaptive band queries. For the random targets in our construction, an estimate at a given noise level usually reveals nothing beyond a common background except over a restricted range of target sizes. We express these ranges on a common logarithmic scale. Neighboring levels have overlapping ranges. We bound their combined coverage and use that bound to control the entire sequence of adaptive queries. Figure 1 illustrates the connection between a band of denoising estimates and its shared coverage. Our results give three concrete consequences:

- **An additive bound for a whole band.** Exact Hamming geometry bounds the combined coverage by the displacement of the ranges plus one range width. The resulting charge is the band half width plus one logarithmic term, rather than one such term for each returned estimate (Section 4).

- **A lower bound that allows adaptive requests.** For each declared evaluation cap, we construct a hard distribution before the algorithm is chosen. Successful approximate sampling requires a worst case radial budget linear in dimension, even when query points, band widths, and stopping times adapt to previous responses (Theorem 2).

- **Separate bounds on requests and evaluations.** At fixed accuracy and moment order, the total number of estimates must scale at least as dimension divided by its logarithm on our hard distributions. For widths whose cap condition permits at least one query, the fixed width result identifies when increasing the width lowers the query bound while the evaluation bound remains linear (Corollaries 3 and 4).

Random codebook constructions and comparisons of response histories already appear in Xun & Price (2026). Our new step is to control the overlap of a contiguous band and turn it into an additive budget for adaptive sampling. This separates reductions in the number of requests from the individual denoising evaluations they contain. The budget theorem, fixed width result, and core lemmas have formal proofs in the Lean 4 proof assistant.

2 MODEL AND COST

Fix a dimension $d \in \mathbb{N}$ and let $V_d = \{-1, +1\}^d$ be the binary hypercube. For $x, y \in V_d$, their Hamming distance $\mathrm{Ham}(x, y)$ counts differing coordinates. The Hamming shell of radius k around x and its cardinality are

$$H_k(x) = \{y \in V_d : \mathrm{Ham}(x, y) = k\}, \qquad B_k = |H_k(x)| = \binom{d}{k}.$$

All logarithms are natural. Radii, band centers, and band half widths belong to $\{0, \dots, d\}$.

A random codebook of size n is $S = (Y_1, \dots, Y_n)$ with independent $Y_i \sim \mathrm{Unif}(V_d)$. The sampling target is its empirical measure $\nu_S = n^{-1} \sum_i \delta_{Y_i}$. We write $\rho(n) = d^{-1} \log n$ for the codebook rate. The lower bound hides this rate from the sampler, which accesses the target through the denoising interface.

At radius k, the fixed energy channel chooses a uniform codeword and flips exactly k uniformly selected coordinates. The squared Euclidean distance between input and output is thus $4k$. Its output distribution is denoted $P_{S,k}$. For an observed point x, define the shell occupancy

$$N_{S,k}(x) = \sum_{i=1}^{n} \mathbf{1}\{Y_i \in H_k(x)\}.$$

Then $P_{S,k}(x) = N_{S,k}(x)/(nB_k)$. When $N_{S,k}(x) > 0$, the posterior mean denoiser, the conditional mean of the original codeword given x, is

$$m_{S,k}(x) = \frac{1}{N_{S,k}(x)} \sum_{i : Y_i \in H_k(x)} Y_i.$$

At points of zero channel mass we set $m_{S,k}(x)$ to the null denoiser under $U = \mathrm{Unif}(V_d)$,

$$m_{U,k}(x) = \Big(1 - \frac{2k}{d}\Big) x. \tag{1}$$

The occupancy has law $N_{S,k}(x) \sim \mathrm{Bin}(n, B_k/2^d)$ and mean $\lambda(n, k) = nB_k/2^d$. Define the window anchor

$$a_k = \log 2 - \frac{1}{d} \log B_k. \tag{2}$$

The exact identity

$$\log \lambda(n, k) = d\big(\rho(n) - a_k\big) \tag{3}$$

places expected occupancy one at $\rho(n) = a_k$. Increasing the codebook rate increases occupancy exponentially. Section 4.1 locates the informative window between sparse and dense occupancy thresholds.

The sampler receives approximate shell denoisers. Fix a moment order $p \geq 2$ and an accuracy level $\varepsilon \in (0, 1)$.

Definition 1 (Shellwise normalized accuracy)**.** A response family $\{\widehat{m}_{S,k}\}_{k=0}^{d}$ is (p, ε) accurate when every codebook S and every radius k satisfy

$$\mathbb{E}_{X \sim P_{S,k}} \left(\frac{\|\widehat{m}_{S,k}(X) - m_{S,k}(X)\|_2}{\sqrt{d}} \right)^p \leq \varepsilon^p. \tag{4}$$

The expectation includes any randomness in the returned estimate.

Our bounds concern access to separate fixed energy shell means. A Gaussian denoiser combines radii through posterior mixture weights. The ordered list specified below does not include these weights, so the bounds do not establish the same evaluation cost for Gaussian denoising.

A band query chooses a point x, a center radius k_0, and a half width b. Its clipped radius set is

$$K_b(k_0) = \{k \in \{0, \dots, d\} : |k - k_0| \leq b\}, \tag{5}$$

and the interface returns the ordered list

$$\mathsf{List}_{S,k_0,b}(x) = \big(\widehat{m}_{S,k}(x)\big)_{k\in K_b(k_0)}.$$

The algorithm chooses each query from the preceding transcript and may stop at a transcript dependent time. Write (x_t, k_t, b_t) for the query at time t.

Three costs describe one transcript. The number of list calls is Q. The total number of returned shell entries is

$$R = \sum_{t=1}^{Q} |K_{b_t}(k_t)|. \tag{6}$$

For a declared cap $R_\star \geq 1$, define

$$\Lambda_\star = 1 + \log d + \log(1 + R_\star) + p\log(1/\varepsilon) \tag{7}$$

and the associated radial budget

$$\mathcal{B}_\star = \sum_{t=1}^{Q} (b_t + \Lambda_\star). \tag{8}$$

Thus Q counts interactions, R counts returned shell denoisers, and $\mathcal{B}_\star$ charges an upper bound on the candidate rate coverage of each call under the declared shell cap. The terms are accumulated over calls, including repeated coverage. The dependence on $R_\star$ and ε enters through the informative windows and the probability of departures from the background response in Section 4.1.

The hard distribution chooses the codebook rate from a fixed compact interval. Fix $0 < q_- < q_+ < 1/2$. Let $h(q) = -q\log q - (1-q)\log(1-q)$ be the binary entropy and set $a(q) = \log 2 - h(q)$. We choose a compact interval $I = [\kappa_L, \kappa_U]$ with $\kappa_L < \kappa_U$ inside the interior of $a([q_-, q_+])$. Hamming symmetry gives $a_k = a_{d-k}$, so every hidden rate in I corresponds to a lower branch below $d/2$ and its reflected upper branch. The proof uses both branches when a band crosses the midpoint.

3 Lower Bounds for Adaptive Band Queries

For each declared shell cap, we construct a distribution of codebooks and accurate denoising response rules on which a small radial budget forces large total variation error. The distribution is fixed before the algorithm is chosen.

Theorem 2 (Adaptive radial budget law)**.** *Fix $p \geq 2$ and the hidden rate interval I from Section 2. There are constants $c_{\mathrm{bud}} > 0$ and d_0, depending only on p and I, with the following property. Let $d \geq d_0$, $\varepsilon \in (0, 1/4)$, and let $R_\star \geq 1$ be an integer shell entry cap. Set*

$$\Lambda_\star = 1 + \log d + \log(1 + R_\star) + p\log(1/\varepsilon).$$

There is a distribution $\mathcal{D}_{d,R_\star,\varepsilon}$ over a hidden rate, a random codebook S, and a (p,ε) accurate response family. This distribution is fixed before the horizon, state space, and algorithm. Every randomized adaptive algorithm whose possible transcript paths satisfy

$$\underbrace{\sum_t |K_{b_t}(k_t)| \leq R_\star}_{\textit{shell entries}}, \qquad \underbrace{\sum_t (b_t + \Lambda_\star) \leq c_{\mathrm{bud}} d}_{\textit{radial budget}} \tag{9}$$

produces an output $\widehat{X}$ with

$$d_{\mathrm{TV}}\big(\mathcal{L}(\widehat{X} \mid S), \nu_S\big) \geq 0.9$$

on a set of instances with $\mathcal{D}_{d,R_\star,\varepsilon}$ probability at least 0.9. Here $\mathcal{L}(\widehat{X} \mid S)$ is the conditional output law and d_{TV} is total variation distance.

Under the declared shell cap, total variation error below 0.9 on more than 0.1 of these instances requires some possible transcript with radial budget above $c_{\mathrm{bud}} d$, and therefore

$$\sum_t \big(b_t + \Lambda_\star\big) = \Omega(d). \tag{10}$$

Centers, half widths, and stopping times may all depend on previous responses. The conclusion concerns the worst case transcript budget. When $\Lambda_\star > c_{\mathrm{bud}} d$, the budget premise permits zero active queries, and the proof uses support separation alone (Appendix F).

Each call returns at least its half width plus one shell entry. Bounding the radial charge in terms of the total entry count gives an evaluation lower bound uniform over adaptive widths. Each declared cap $R_\star$ indexes its own hard distribution.

Corollary 3 (Uniform shell evaluation lower bound)**.** *Let c_{bud}, d_0 be as in Theorem 2. Fix $d \geq d_0$, $\varepsilon \in (0, 1/4)$, and an integer $R_\star \geq 1$. Suppose an algorithm performs at most $R_\star$ shell evaluations on every path and achieves total variation error below 0.9 with $\mathcal{D}_{d,R_\star,\varepsilon}$ probability greater than 0.1. Then*

$$R_\star(1 + \Lambda_\star) > c_{\mathrm{bud}} d, \qquad R_\star = \Omega\left(\frac{d}{\log d + p \log(1/\varepsilon)}\right). \tag{11}$$

For constant p and $\varepsilon \geq d^{-O(1)}$, each proposed shell budget below a sufficiently small constant multiple of $d/\log d$ admits a hard distribution under which every algorithm respecting that budget incurs total variation error at least 0.9 on at least 0.9 of the instances.

Proof. Every clipped band contains at least $b_t + 1$ radii. If a transcript uses Q calls and at most $R_\star$ shell entries, then $\sum_t b_t \leq R_\star - Q$ and $Q \leq R_\star$. Hence

$$\sum_t (b_t + \Lambda_\star) = \sum_t b_t + Q\Lambda_\star \leq R_\star(1 + \Lambda_\star).$$

Theorem 2 forces the first inequality in (11). If $R_\star \leq d$, then $\Lambda_\star \leq 2 + 2\log(1 + d) + p \log(1/\varepsilon)$, which gives the displayed asymptotic bound. The case $R_\star > d$ is immediate. □

For a fixed half width, the same budget yields separate bounds on calls and shell evaluations.

Corollary 4 (Fixed width query lower bound)**.** *Fix $p \geq 2$. There are constants $c_{\mathrm{fix}} > 0$ and d_0 such that for every $d \geq d_0$, $\varepsilon \in (0, 1/4)$, and $b \in \{0, \ldots, d\}$, define*

$$\Lambda_b = 1 + \log d + \log(1 + b) + p \log(1/\varepsilon). \tag{12}$$

For every integer cap $Q_\star \geq 1$ satisfying

$$Q_\star \leq c_{\mathrm{fix}} \frac{d}{b + \Lambda_b}, \tag{13}$$

there is a hard distribution $\mathcal{D}_{d,b,Q_\star,\varepsilon}$, fixed before the horizon, state space, and algorithm, under which every adaptive algorithm with at most $Q_\star$ width b queries incurs total variation error at least 0.9 on at least 0.9 of the instances. Equivalently, the fixed width query lower bound scales as

$$Q = \Omega\left(\frac{d}{b + \Lambda_b}\right).$$

When $c_{\mathrm{fix}} d/(b + \Lambda_b) \geq 1$, the cap condition admits a positive integer query budget. In this regime, each width b query returns at least $b + 1$ shell entries, giving

$$R = \Omega\left(\frac{d(b + 1)}{b + \Lambda_b}\right). \tag{14}$$

Let $\Lambda = 1 + \log d + p \log(1/\varepsilon)$. Then $\Lambda \leq \Lambda_b \leq 2\Lambda$ for $0 \leq b \leq d$. Figure 2 shows the lower bound scales across three fixed width regimes. For bounded b, the query and shell lower bounds are both of order d/Λ. For $1 \ll b \ll \Lambda$, their scales separate as d/Λ and $d(b + 1)/\Lambda$, respectively. At $b \asymp \Lambda$, the shell lower bound reaches linear order in d, while the query lower bound remains of order d/Λ. For larger widths, the query lower bound scales as d/b and the shell lower bound stays linear.

Theorem 5 (Two sided Hamming anchor drift)**.** *Fix $0 < q_0 < q_1 < 1/2$. There are constants $0 < c_{\mathrm{drift}} < C_{\mathrm{drift}} < \infty$ and d_1 such that for every $d \geq d_1$ and all $k, \ell \in \{0, \ldots, d\}$ with $k/d, \ell/d \in [q_0, q_1]$,*

$$c_{\mathrm{drift}} \frac{|k - \ell|}{d} \leq |a_k - a_\ell| \leq C_{\mathrm{drift}} \frac{|k - \ell|}{d}.$$

The exact binomial ratio in (16) gives the adjacent anchor increment. Summing these increments proves Theorem 5 (Appendix D) and bounds the anchor span of a band on a compact branch in proportion to its half width divided by dimension.

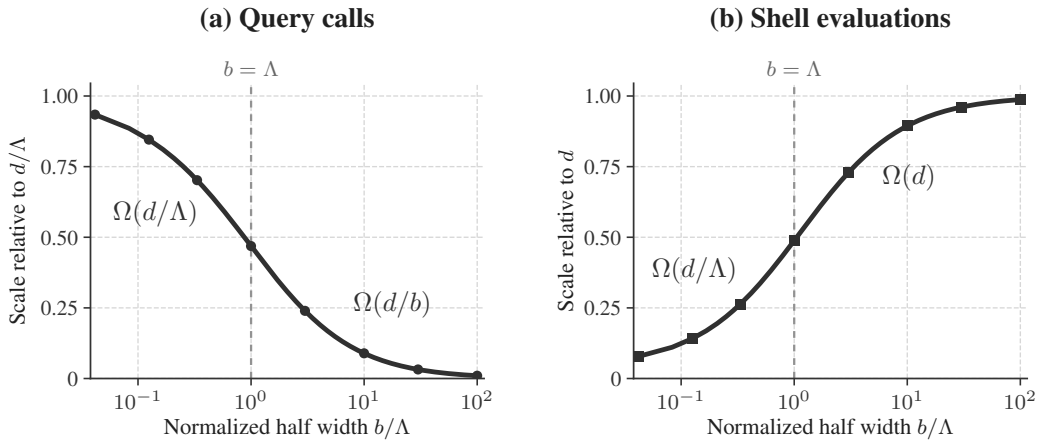


Figure 2: Algebraic lower bound scales for fixed width queries. Panel (a) plots $\Lambda/(b+\Lambda_b)$, the query bound scale in units of d/Λ. Panel (b) plots $(b+1)/(b+\Lambda_b)$, the shell bound scale in units of d. Curves connect integer half widths for $d = 10^6$, $p = 2$, and $\varepsilon = 10^{-2}$, with $\Lambda_b = \Lambda + \log(1+b)$. Dashed vertical lines mark $b = \Lambda$. Theorem constants are suppressed. Corollary 4 applies where its cap condition admits a positive integer query budget.

4 Why Wide Bands Have Redundant Coverage

The proof bounds the union of informative windows along a null transcript independent of the target. A small union and a small total disagreement probability outside the windows keep the real and null responses likely to agree. Independence prevents the null output from sampling the hidden codebook accurately.

4.1 From occupancy to informative windows

Fix a constant $\gamma > 0$. Lemma 17 gives a candidate grid $G \subset I$ of $\Theta(d)$ realizable rates, with adjacent gaps in $[\gamma/d, 2\gamma/d]$. The hard distribution draws $\kappa \sim \mathrm{Unif}(G)$, sets $n = e^{d\kappa}$, and samples n uniform codewords. The response rule in Appendix C is fixed for this instance before the algorithm is chosen.

At a fixed query point, a sparse shell is usually empty. In a dense shell, averaging many random codewords brings the posterior mean close to the uniform background mean. On dense shells, the response rule returns the background mean when a channel averaged squared error test certifies the pth moment accuracy in Definition 1, and the exact posterior otherwise. Lemmas 10 and 11 establish accuracy for every codebook and a small probability of disagreement with the background outside an intermediate occupancy range.

For the shell cap $R_\star$, set $\delta_0 = c_0/R_\star$ with a fixed small constant c_0, and let

$$\lambda_- = \delta_0, \qquad \lambda_+ = \frac{1}{\delta_0 2^{2-p}\varepsilon^p}.$$

Since $\log \lambda(n,k) = d(\kappa - a_k)$, the intermediate occupancy range corresponds to the informative rate window

$$J(k) = \left[a_k + \frac{1}{d}\log\lambda_-, a_k + \frac{1}{d}\log\lambda_+\right]. \tag{15}$$

Write the logarithmic offsets as $L = d^{-1}\log\lambda_-$ and $U_{\mathrm{off}} = d^{-1}\log\lambda_+$. The windows are common translates $J(k) = [a_k + L, a_k + U_{\mathrm{off}}]$ of width $w = U_{\mathrm{off}} - L \le C_{\mathrm{win}}\Lambda_\star/d$. Outside $J(k)$, the response differs from the null mean with probability at most δ_0 under the original codebook law. The cap $R_\star$ bounds the sum of these disagreement probabilities over all returned shell entries.

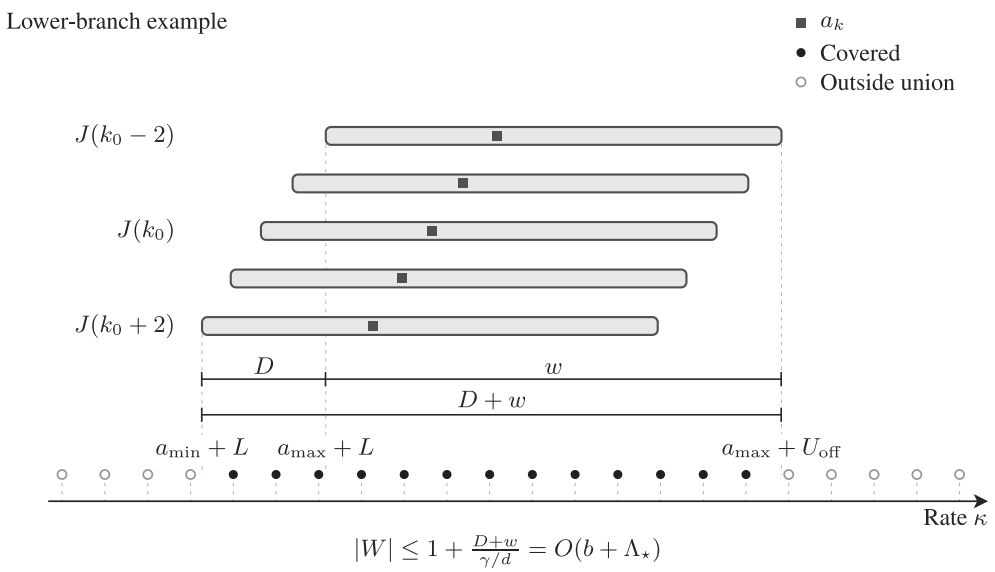


Figure 3: A band pays for the anchor span plus one window width. Five neighboring shells on a lower Hamming branch give the shaded windows $J(k)$, with filled squares marking their anchors a_k. The three endpoint guides join $D = a_{\max} - a_{\min}$ and $w = U_{\text{off}} - L$ at $a_{\max} + L$, giving an enclosing interval of length $D + w$. Filled circles are the 13 distinct covered rates W among the 22 illustrated grid points, while open circles lie outside the window union. Dashed lines align equal rate coordinates. Minimum grid separation γ/d converts the enclosing length into the displayed additive count, including the endpoint term 1.

4.2 Additive coverage from contiguous shells

The windows of a contiguous band fit inside one interval whose length is the anchor span plus one window width. On a compact Hamming branch, the exact identity

$$a_{k+1} - a_k = -\frac{1}{d} \log \frac{d-k}{k+1} \tag{16}$$

gives adjacent anchor drift of order $1/d$. For a band $K_b(k_0)$, let $a_{\min} = \min_{k \in K_b(k_0)} a_k$ and $a_{\max} = \max_{k \in K_b(k_0)} a_k$. Their span $D = a_{\max} - a_{\min}$ is $O(b/d)$. The windows lie inside $[a_{\min} + L, a_{\max} + U_{\text{off}}]$, an interval of length $D + w$. Figure 3 splits this length at $a_{\max} + L$ into the anchor span D and one window width w. For the covered candidates $W = G \cap \bigcup_{k \in K_b(k_0)} J(k)$, the minimum grid separation γ/d therefore gives

$$|W| \le 1 + \frac{D+w}{\gamma/d} = O(b + \Lambda_\star).$$

Summing the individual shell coverage bounds would count overlapping regions repeatedly, charging $O((b+1)\Lambda_\star)$ candidates. Bounding the enclosing interval reduces this charge to $O(b + \Lambda_\star)$, with one copy of the window width term per call.

For arbitrary bands, Lemma 15 localizes the radii whose windows meet I to a compact lower branch and its reflection. On the interval enclosing both branches, an absolute drift bound controls the span of contributing anchors even when a band crosses $d/2$. Together with $a_k = a_{d-k}$ and endpoint clipping, Lemma 16 gives the same additive count. The separation conditions hold whenever the radial budget permits an active query (Appendix F).

4.3 Coverage along an independent null path

To handle adaptive choices, run the algorithm with $m_{U,k}(x)$ as the response at every shell. This completed null transcript is independent of the hidden rate and codebook. Along that path, let

Table 1: Query responses and resource accounting in two random codebook constructions. Xun & Price (2026) study one Gaussian smoothed score per call. This paper studies an adaptive band of fixed energy denoisers and charges its accumulated hidden rate coverage. A score is the gradient of a log density. The bounds in the last column have the hypotheses of Theorem 2 and Corollary 3.

	Xun & Price (2026)	This paper
Response	one Gaussian smoothed score at a selected scale	a list of fixed energy denoisers on an adaptive radius band
Adaptive choice	one scale per query	center, half width, and stopping time
Charged resource	query count Q	radial budget $\mathcal{B}_\star = \sum_t (b_t + \Lambda_\star)$
Geometric unit	one informative scale window	overlap among contiguous shell windows
Main consequence	$\sqrt{d}$ Gaussian query lower bound up to logarithmic factors	$\mathcal{B}_\star = \Omega(d)$ and $R = \Omega(d/(\log d + p\log(1/\varepsilon)))$

(x_t^0, k_t^0, b_t^0) be the queries and define

$$W_t^0 = G \cap \bigcup_{k \in K_{b_t^0}(k_t^0)} J(k), \qquad \mathsf{Hit} = \left\{ \kappa \in \bigcup_t W_t^0 \right\}.$$

Conditioning on the null path fixes its windows while leaving κ uniform on G. Consequently,

$$\mathbb{P}(\mathsf{Hit} \mid H^0) \le \frac{\sum_t |W_t^0|}{|G|} \le \frac{C \sum_t (b_t^0 + \Lambda_\star)}{|G|},$$

where H^0 denotes the completed null transcript. The pathwise radial budget bounds this probability, including for policies that adapt their centers, widths, and stopping times.

4.4 From joint responses to sampling error

At a common transcript prefix, couple the entire real and null response lists using the same realized query. Matching individual response distributions does not control their dependence: the uniform laws on $\{00, 11\}$ and $\{01, 10\}$ have identical coordinate marginals but disjoint joint supports. Our response rule is deterministic given the codebook. On the common codebook probability space, a list disagreement requires at least one component disagreement. At a fixed query under the original codebook law, a hidden rate outside every window in the band gives a disagreement probability at most $|K_{b_t}(k_t)|\delta_0$.

Let τ be the first response disagreement. At each common prefix, the proof drops the earlier agreement indicator before averaging the mismatch charge under the original codebook law. This preserves the estimate at each fixed query point despite adaptive query selection. Summing the charges gives at most $R_\star \delta_0 = c_0$ in addition to the hit probability. Appendix E constructs the coupling using probability kernels, which assign each history a joint response distribution that varies measurably with that history.

On transcript agreement, the two runs use the same output randomness and produce the same output. The null output is independent of the codebook, whereas the target is supported on its codewords. Any fixed output point belongs to this random support with exponentially small probability. Support separation therefore converts the hit and disagreement bounds into the total variation lower bound. Appendix F combines these estimates with the constants in Theorem 2.

5 Related Work

Lower bounds for diffusion sampling quantify the number of queries needed to sample an unknown target. Xun & Price (2026) proved a $\sqrt{d}$ lower bound up to logarithmic factors for Gaussian smoothed

scores. Their proof uses random codebook hard instances, a packing of candidate rates, an instance independent null transcript, coupling at the first response disagreement, and separation of output supports. Their spherical noise warmup illustrates the single level intuition behind informative scale windows.

Our contribution is an additive coverage bound and radial information budget for adaptive contiguous bands. Exact Hamming geometry controls the union of neighboring informative windows by the anchor span plus one window width. For each declared evaluation cap, the resulting hard distribution forces the shell evaluation lower bound in Corollary 3, allowing adaptive width choices. Table 1 compares the two access models and their charged resources.

Other lower bounds study access to stochastic gradient estimates (Chatterji et al., 2020). For Gaussian targets, Liu (2026) shows that exact smoothed score queries give access to resolvents, the inverses of shifted precision matrices. Here a precision matrix is an inverse covariance matrix, and a query reveals the action of a resolvent on a vector.

Upper bound analyses study concrete samplers under accurate scores, including reverse stochastic differential equation methods (Chen et al., 2023), stochastic localization (Benton et al., 2024), the higher order and exponential integrators of Lu et al. (2022) and Karras et al. (2022), and smooth regimes with sublinear iteration counts (Zhang et al., 2025).

In the fixed energy model, the translated shell response $u_k(x) = m_{S,k}(x) - x$ gives an equivalent response representation. Subtracting the known query point x from each approximate response preserves its error, $\widehat{u}_k(x) - u_k(x) = \widehat{m}_{S,k}(x) - m_{S,k}(x)$. The inverse translation recovers the denoiser at the same query and shell cost, so the lower bounds apply to both representations.

6 Conclusion

At fixed denoising accuracy and moment order, our hard distributions in the fixed energy band model give a worst case shell evaluation lower bound proportional to dimension divided by its logarithm. The bound allows query points, band widths, and stopping times to adapt to earlier responses. Exact Hamming geometry bounds the combined coverage of neighboring shells by their anchor span plus one window width. Coupling complete response histories turns that overlap bound into the radial information budget. A call that returns several neighboring denoisers can therefore expose overlapping rate windows. Counting the call alone conceals the individual evaluations needed to obtain those responses.

AI use statement

AI tools assisted with wording and proofreading of the manuscript.

Reproducibility statement

Section 2 and Appendix C specify the model, response rule, and constants used by the main theorem. The appendices give the complete elementary arguments and the measurable joint kernel coupling. Theorem 2, Corollary 4, Theorem 5, and their core lemmas have formal proofs in Lean 4 on top of the mathlib mathematics library. Appendix H gives the exact statement correspondence. The supplementary material includes the formal development, pinned toolchain, and a log of rechecking proof declarations with Lean's proof checking kernel.

References

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.

Niladri S. Chatterji, Peter L. Bartlett, and Philip M. Long. Oracle lower bounds for stochastic gradient sampling algorithms. *arXiv preprint arXiv:2002.00291*, 2020.

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.

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.

Jingbo Liu. Smoothed score queries and the complexity of sampling. *arXiv preprint arXiv:2605.27769*, 2026.

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.

Zhiyang Xun and Eric Price. Query lower bounds for diffusion sampling. *arXiv preprint arXiv:2604.10857*, 2026.

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 Notation and Quantifier Conventions

Table 2: Notation for the fixed energy model and radial budget.

Symbol	Meaning
$V_d = \{-1,+1\}^d$, Ham	hypercube and Hamming distance
$H_k(x)$, $B_k = \binom{d}{k}$	shell of radius k and its cardinality
$S = (Y_1, \dots, Y_n)$, ν_S	random codebook and its empirical target
$\rho(n) = \frac{1}{d}\log n$	rate, with hidden interval $I = [\kappa_L, \kappa_U] \subset (0, \log 2)$
$P_{S,k}$, $m_{S,k}$, $m_{U,k}$	fixed energy channel, posterior mean, null mean
$N_{S,k}(x)$, $\lambda(n,k) = nB_k/2^d$	shell occupancy and its mean
$a_k = \log 2 - \frac{1}{d}\log B_k$	window anchor of shell k on the rate axis
$K_b(k_0)$	clipped band $\{k \in \{0,\dots,d\} : \lvert k - k_0\rvert \le b\}$
Q, R	query count and total returned shell entries, (6)
$\Lambda_\star$	logarithmic term $1 + \log d + \log(1+R_\star) + p\log(1/\varepsilon)$
$\mathcal{B}_\star$	radial budget $\sum_t (b_t + \Lambda_\star)$, (8)
δ_0	bound on response disagreement outside the window, (17)
$\lambda_\pm$	occupancy thresholds, (17)
θ_p	channel averaged squared error threshold, (17)
$J(k)$	informative rate window of shell k, (15)
G	candidate rate grid, spacing $\ge \gamma/d$

All logarithms are natural. Radii, band centers, and half widths are integers in $\{0,\dots,d\}$. Constants named $c, C, c_0, C_0, \dots$ depend only on p and the fixed interval I (equivalently on q_-, q_+, γ) unless stated otherwise, and never on $d, \varepsilon, R_\star, b$, or the algorithm.

The order of quantifiers matters throughout the paper. Every lower bound has the form

$$\forall \text{ budget parameters} \quad \exists \text{ hard distribution } \mathcal{D} \quad \forall \text{ horizon, state space, algorithm.}$$

The hard distribution may depend on d, p, ε, and the declared budget ($R_\star$ in Theorem 2, $Q_\star$ and b in Corollary 4), and never on the algorithm, its horizon, its state space, or its realized transcript. Budget hypotheses are imposed on every possible transcript path.

We use probability kernels, which assign a probability distribution to each input, measurably in that input. All state, history, query, joint response, and output spaces are standard Borel: their measurable sets are the Borel sets of some complete separable metric. An algorithm is a randomized policy on

the state space. At each step it either stops or emits a query (x_t, k_t, b_t) from the current transcript and its independent randomness. It then receives $\mathsf{List}_{S,k_t,b_t}(x_t)$ and eventually emits $\widehat{X}$ through a measurable output kernel. A stopped run remains in a cemetery state, with an empty window and zero charge.

Figure 1 depicts the band response and its analytical rate windows using the same orthographic viewing directions. In the shell view, spatial radii, state positions, and point counts are schematic. The blue layers select three complete shells through a shared octant cut. Dashed latitude arcs indicate portions occluded by retained shell surfaces. In the proof view, equal width windows shift left along the lower Hamming branch as the shell index increases. Height and depth separate the windows for display. Guides project their corners to the rate plane, and dashed traces mark upper short edges projected onto lower bands. Hatching identifies coverage by higher bands in a top view. Darker floor regions indicate greater overlap, rather than probability or information strength. Grey candidates lie outside the union. Their possible effect is controlled by the bound on departures from the background response.

B Estimates for One Shell

Lemma 6 (Sparse occupancy)**.** *If $N \sim \mathrm{Bin}(n, u)$ with success probability u and mean $\lambda = nu$, then $\mathbb{P}[N > 0] \le \lambda$. In particular $\mathbb{P}_S[N_{S,k}(x) > 0] \le \lambda(n, k)$ for every x and k.*

Proof. $\mathbb{P}[N \ge 1] \le \mathbb{E}N = \lambda$ by Markov's inequality. □

Lemma 7 (Within shell covariance trace)**.** *If $Y \sim \mathrm{Unif}(H_k(x))$ then $\operatorname{tr}\operatorname{Cov}(Y) = 4k(1-k/d) \le d$.*

Proof. Coordinate j satisfies $\mathbb{P}[Y_j = -x_j] = k/d$, hence $\mathbb{E}Y_j = (1 - 2k/d)x_j$ and $\mathrm{Var}(Y_j) = 1 - (1 - 2k/d)^2$. The sum over j gives $d - d(1 - 2k/d)^2 = 4k(1 - k/d)$, which is maximized at $k = d/2$ with value d. □

Lemma 8 (Channel averaged deviation on dense shells)**.** *Let $J_k(S) = \mathbb{E}_{X\sim P_{S,k}}\|m_{S,k}(X) - m_{U,k}(X)\|_2^2$. Then $\mathbb{E}_S J_k(S) \le d/\lambda(n, k)$.*

Proof. Write $J_k(S) = \sum_x \frac{N_{S,k}(x)}{nB_k}\|m_{S,k}(x) - m_{U,k}(x)\|_2^2$. Fix x and put $N = N_{S,k}(x)$. Conditionally on $N = r \ge 1$ the codewords in $H_k(x)$ are independent and uniform on the shell, so by Lemma 7

$$\mathbb{E}\Big[\big\|\tfrac{1}{r}\textstyle\sum_{i\le r} Z_i - \mathbb{E}Z_1\big\|_2^2 \;\Big|\; N = r\Big] = \tfrac{1}{r}\operatorname{tr}\operatorname{Cov}(Z_1) \le \tfrac{d}{r}.$$

The factor $N = r$ reduces this bound to d. The expected contribution of x is therefore at most $d/(nB_k)$. The sum over all 2^d points is $2^d d/(nB_k) = d/\lambda$. □

Lemma 9 (Dense shell deviation at a fixed query point)**.** *With the convention $m_{S,k}(x) = m_{U,k}(x)$ on $\{N_{S,k}(x) = 0\}$,*

$$\mathbb{E}_S\|m_{S,k}(x) - m_{U,k}(x)\|_2^2 \le \frac{2d}{\lambda(n, k)} \qquad \textit{for every fixed } x, k.$$

Proof. Conditionally on $N = r \ge 1$ the mean square deviation is at most d/r, and $1/r \le 2/(r + 1)$ for $r \ge 1$. For $N \sim \mathrm{Bin}(n, u)$,

$$\mathbb{E}\frac{1}{N+1} = \int_0^1 \mathbb{E}[t^N]\,\mathrm{d}t = \int_0^1 (1 - u + ut)^n\,\mathrm{d}t = \frac{1 - (1-u)^{n+1}}{(n+1)u} \le \frac{1}{nu} = \frac{1}{\lambda}.$$

Hence $\mathbb{E}[\mathbf{1}_{N\ge1}/N] \le 2/\lambda$. □

C Shell Responses for the Lower Bound

Fix a total shell entry cap $R_\star$, let c_0 be a small absolute constant, and set

$$\delta_0 = \frac{c_0}{R_\star}, \qquad \lambda_- = \delta_0, \qquad \theta_p = 2^{2-p}\varepsilon^p d, \qquad \lambda_+ = \frac{1}{\delta_0\, 2^{2-p}\varepsilon^p}. \tag{17}$$

For every codebook S and radius k, with $\lambda = \lambda(n,k)$, define $\widehat{m}_{S,k}$ by:

1. If $\lambda \le \lambda_-$, return $m_{S,k}(x)$ when $N_{S,k}(x) > 0$, and $m_{U,k}(x)$ otherwise.
2. If $\lambda_- < \lambda < \lambda_+$, return the true posterior mean.
3. If $\lambda \ge \lambda_+$ and $J_k(S) \le \theta_p$, return $m_{U,k}(x)$ for every x.
4. If $\lambda \ge \lambda_+$ and $J_k(S) > \theta_p$, return the true posterior mean.

The response rule is deterministic given S, and the third and fourth cases depend on S only through the scalar $J_k(S)$.

Lemma 10 (Accuracy of the constructed responses)**.** *For every $p \ge 2$, $\varepsilon \in (0,1)$, S and k, the family above satisfies Definition 1 with parameter ε.*

Proof. Under $X \sim P_{S,k}$ we have $N_{S,k}(X) > 0$ almost surely, so in case (1) the error is identically zero. Cases (2) and (4) also return the exact posterior. In case (3) put $Z = \|m_{U,k}(X) - m_{S,k}(X)\|_2/\sqrt{d}$. Both means lie in $[-1,1]^d$, so $0 \le Z \le 2$, and $p \ge 2$ gives $Z^p \le 2^{p-2}Z^2$. Hence

$$\mathbb{E}Z^p \le 2^{p-2}\,\frac{J_k(S)}{d} \le 2^{p-2}\,2^{2-p}\varepsilon^p = \varepsilon^p. \qquad \square$$

Lemma 11 (Agreement at a fixed query point off the window)**.** *Fix x and k. If $\lambda(n,k) \notin [\lambda_-, \lambda_+]$ then*

$$\mathbb{P}_S\big[\widehat{m}_{S,k}(x) \ne m_{U,k}(x)\big] \le \delta_0.$$

Proof. If $\lambda \le \lambda_-$ then disagreement requires $N_{S,k}(x) > 0$, which has probability at most $\lambda \le \delta_0$ by Lemma 6. If $\lambda \ge \lambda_+$ then disagreement requires $J_k(S) > \theta_p$, and by Lemma 8 and Markov's inequality,

$$\mathbb{P}[J_k(S) > \theta_p] \le \frac{d/\lambda}{2^{2-p}\varepsilon^p d} \le \frac{1}{\lambda_+ 2^{2-p}\varepsilon^p} = \delta_0. \qquad \square$$

The probability in Lemma 11 is taken under the *original* codebook law, before conditioning on the algorithm's history. Section 4.4 uses this estimate by dropping the earlier agreement indicator before averaging the response mismatch charge.

D Exact Hamming Geometry

D.1 Windows, anchors, and drift

By (2) the intermediate occupancy condition $\lambda_- \le \lambda(n,k) \le \lambda_+$ is equivalent to $\kappa \in J(k)$ with $J(k)$ as in (15), whose width is exactly

$$|J(k)| = \frac{1}{d}\left(2\log\frac{1}{\delta_0} + p\log\frac{1}{\varepsilon} + (p-2)\log 2\right) \;\le\; C\,\frac{\Lambda_\star}{d} \tag{18}$$

when $\delta_0 = c_0/R_\star$. The width is independent of k, so all windows are translates of the same interval.

Lemma 12 (Adjacent anchors)**.** *For $0 \le k < d$, $a_{k+1} - a_k = -\frac{1}{d}\log\frac{d-k}{k+1}$.*

Proof. Immediate from $\binom{d}{k+1}/\binom{d}{k} = (d-k)/(k+1)$ and (2). $\square$

Proof of Theorem 5. If $j/d \in [q_0, q_1]$ with $q_1 < 1/2$ then for all large d the ratio $(d-j)/(j+1)$ lies in a fixed compact subinterval of $(1, \infty)$, so $0 < c_{\mathrm{loc}} \le \log \frac{d-j}{j+1} \le C_{\mathrm{loc}} < \infty$. A sum of Lemma 12 over $j = k, \dots, \ell - 1$ gives $c_{\mathrm{loc}}|k-\ell|/d \le |a_k - a_\ell| \le C_{\mathrm{loc}}|k-\ell|/d$. No Stirling estimate is used, so the proof uses the exact increment and avoids the $O(\log d/d)$ error from Stirling's approximation. □

Proposition 13 (Diameter of a branch contained band window)**.** *Fix $0 < q_0 < q_1 < 1/2$ and $0 < c_H \le C_H$. For all large d, let $H > 0$, $b \in \mathbb{N}$, and k_0 be such that every $k \in K_b(k_0)$ satisfies $k/d \in [q_0, q_1]$, and let $J(k) = a_k + [L_d, U_d]$ with $c_H H/d \le U_d - L_d \le C_H H/d$. Then there are constants $0 < c' < C'$ depending only on q_0, q_1, c_H, C_H with*

$$c' \frac{b+H}{d} \;\le\; \operatorname{diam}\Big(\bigcup_{k \in K_b(k_0)} J(k) \Big) \;\le\; C' \frac{b+H}{d}.$$

Proof. The band is unclipped, so its extreme indices differ by $2b$ and, the windows being common translates, $\operatorname{diam}(\bigcup_k J(k)) = \max_k a_k - \min_k a_k + (U_d - L_d)$. Apply Theorem 5 to the two extreme anchors and add the assumed bounds on $U_d - L_d$. □

D.2 The two Hamming branches

Exact symmetry $\binom{d}{d-k} = \binom{d}{k}$ gives $a_k = a_{d-k}$, hence $J(k) = J(d-k)$. Relevant radii therefore occupy a compact lower branch together with its reflected upper branch. Fix $0 < q_L < q_U < 1/2$, occupancy thresholds $0 < L < U$, and a margin $m > 0$ satisfying

$$\kappa_U + m \le a(q_L), \quad a(q_U) \le \kappa_L - m, \qquad \max\Big\{ -\tfrac{\log L}{d}, \tfrac{\log U + \log(d+1)}{d} \Big\} \le m, \tag{19}$$

where $J_{L,U}(k) = [a_k + \frac{\log L}{d}, a_k + \frac{\log U}{d}]$.

Lemma 14 (Relevant radii, lower branch)**.** *Under* (19)*, if $2k \le d$ and $J_{L,U}(k) \cap I \ne \varnothing$ then $q_L \le k/d \le q_U$.*

Proof. The finite entropy estimate $0 \le a_k - a(k/d) \le \log(d+1)/d$ places $a(k/d)$ within m of every rate in $J_{L,U}(k) \cap I$, so (19) gives $a(q_U) \le a(k/d) \le a(q_L)$. On $[0, 1/2]$ the map $q \mapsto a(q) = \log 2 - h(q)$ is strictly decreasing, and $2k \le d$ puts k/d in that range. □

Lemma 15 (Relevant radii on the two reflected branches)**.** *Under* (19)*, if $J_{L,U}(k) \cap I \ne \varnothing$ then $k/d \in [q_L, q_U] \cup [1-q_U, 1-q_L]$.*

Proof. Fold to $\bar{k} = \min\{k, d-k\}$. Then $2\bar{k} \le d$, and $a_{\bar{k}} = a_k$, $J_{L,U}(\bar{k}) = J_{L,U}(k)$. Apply Lemma 14 to $\bar{k}$ and unfold. □

Lemma 16 (Band hit count, arbitrary bands)**.** *Assume* (19)*. Let $s \in \mathbb{N}$ be positive with $\log(U/L) \le s$ and let $G \subset I$ be finite with pairwise spacing at least γ/d. Then for every center k_0 and every half width b,*

$$\Big| G \cap \bigcup_{k \in K_b(k_0)} J_{L,U}(k) \Big| \;\le\; C_{\gamma, q_L}\,(b+s), \qquad C_{\gamma,q_L} = \Big\lceil \max\Big\{ \tfrac{4\log(1/q_L)}{\gamma}, \tfrac{2}{\gamma} \Big\} \Big\rceil + 1.$$

Proof. By Lemma 15 every radius contributing to the intersection lies in the enclosing interval $[q_L d, (1-q_L)d]$, on which Lemma 12 gives the *absolute* bound $|a_{j+1} - a_j| \le \log(1/q_L)/d$. The proof uses neither monotonicity nor a two sided bound, so the estimate survives a band straddling $d/2$. Two contributing radii in the same clipped band therefore have anchors at distance at most $2b\log(1/q_L)/d$, and adding the common window width $\log(U/L)/d \le s/d$ bounds the diameter of the exposed region inside I. A γ/d-separated subset of an interval of that diameter has at most the displayed cardinality, the additive 1 accounting for the endpoint. □

Lemma 17 (Fine rate grid)**.** *Fix a constant length $I \subset (0, \log 2)$ and $\gamma > 0$. For all large d there is $G_d \subset \{\frac{1}{d}\log n : n \in \mathbb{N}, n \ge 1\} \cap I$ with $|G_d| \ge c_I d$ and adjacent gaps in $[\gamma/d, 2\gamma/d]$.*

Proof. Write $I = [\kappa_L, \kappa_U]$, set $\kappa_j = \kappa_L + \frac{\gamma}{2d} + \frac{3\gamma j}{2d}$ and $n_j = \lfloor e^{d\kappa_j} \rfloor$. For large d we have $e^{d\kappa_j} \geq 2$ and $2e^{-d\kappa_j} \leq \gamma/2$, so with $g_j = \frac{1}{d}\log n_j$ and $\varepsilon_j = \kappa_j - g_j$ the floor estimate gives $0 \leq \varepsilon_j \leq \gamma/(2d)$, whence $g_{j+1} - g_j = \frac{3\gamma}{2d} + \varepsilon_j - \varepsilon_{j+1} \in [\gamma/d, 2\gamma/d]$. The padding in κ_0 gives $g_j \geq \kappa_L$, and $g_j \leq \kappa_j \leq \kappa_U$ whenever $\kappa_j \leq \kappa_U$. The number of such indices is at least $|I|d/(3\gamma)$ once $d \geq \gamma/|I|$. □

D.3 Coverage accounting

Proposition 18 (Finite band coverage budget)**.** *Let G be finite, let $W_1, \ldots, W_Q$ be the rate regions exposed by bands of half widths $b_1, \ldots, b_Q$, and suppose $G \subseteq \bigcup_t W_t$ and $|G \cap W_t| \leq C_{\mathrm{hit}}(b_t + H)$ for every t. Then $|G| \leq C_{\mathrm{hit}} \sum_t (b_t + H)$. In particular, $|G| \leq C_{\mathrm{hit}} Q(b_{\max} + H)$.*

Proof. Every point of G is counted at least once, so $|G| \leq \sum_t |G \cap W_t| \leq C_{\mathrm{hit}} \sum_t (b_t + H)$. □

E The Adaptive Coupling of Joint Kernels

The band response is a list, so the proof must couple the full response law. For example, the uniform laws on $\{00, 11\}$ and $\{01, 10\}$ have identical coordinate marginals. Their joint supports are disjoint. Separate coordinate couplings therefore leave a dependent list uncontrolled. We work with the joint response kernel throughout.

Theorem 19 (Random window budget coupling for joint kernels)**.** *Fix a finite horizon T and let all history, query, joint response and output spaces be standard Borel with measurable policy, response, stopping and output kernels. Let G be finite and nonempty, and let Z be a standard Borel payload space. Associate to each $S \in \mathsf{Z}$ a measurable support set $A(S)$ in the output space, with jointly measurable membership map $(S, x) \mapsto \mathbf{1}\{x \in A(S)\}$. Sample a hidden index $K \sim \mathrm{Unif}(G)$ and then S from its original conditional kernel given K. Write the instance as (K, S). Assume the target kernel satisfies $\nu_{K,S}(A(S)) = 1$ almost surely. Let the null run be a process independent of (K, S). At every matching history h_t^0 before a response, the policy samples one action and query q whose realized value is* shared *by the two sides while they remain matched. Let $W_t(h_t^0, q) \subseteq G$ and the nonnegative charge $\delta_t(h_t^0, q)$ be measurable, with a stop action contributing $\varnothing$ and 0. At every common prefix and query, supply a jointly measurable maximal coupling of the real and null joint response kernels: a joint law with these marginals whose disagreement probability equals their total variation distance. For a completed null transcript H^0, evaluate the windows and charges at its history h_t^0 and query q_t^0: $\widehat{W}_t(H^0) = W_t(h_t^0, q_t^0)$ and $\widehat{\delta}_t(H^0) = \delta_t(h_t^0, q_t^0)$. Fix $\beta, \Delta, \eta \geq 0$ and assume*

1. *$\sum_{t \leq T} |\widehat{W}_t(H^0)| \leq \beta|G|$ on every null path.*
2. *For every fixed prefix h_t^0, query q, and $j \notin W_t(h_t^0, q)$,*
$$\mathbb{E}_{S|K=j}\, d_{\mathrm{TV}}\big(P_{j,S,t}(\cdot \mid h_t^0, q), P_{0,t}(\cdot \mid h_t^0, q)\big) \leq \delta_t(h_t^0, q),$$
the expectation being under the original instance law, not *conditioned on agreement at earlier times.*
3. *$\sum_{t \leq T} \widehat{\delta}_t(H^0) \leq \Delta$ on every null path.*
4. *The null output law Q_0 is independent of (K, S) and every fixed output point x satisfies $\mathbb{P}[x \in A(S)] \leq \eta$.*

Let Diff *denote the event that the two completed transcripts differ. Then the two runs admit a sequential coupling with $\mathbb{P}(\mathsf{Diff}) \leq \beta + \Delta$, and for all $a, b > 0$ with $a + b < 1$,*

$$\mathbb{P}_{K,S}\big[d_{\mathrm{TV}}(Q_{K,S}, \nu_{K,S}) \geq 1 - a - b\big] \;\geq\; 1 - \frac{\beta + \Delta}{a} - \frac{\eta}{b}.$$

For the codebook payload $S = (Y_1, \ldots, Y_n)$, take $A(S) = \{Y_1, \ldots, Y_n\}$. The target ν_S retains the multiplicity of each codeword, while $A(S)$ records its support.

Proof. We construct one probability space carrying the instance, the real run, and the null run. First sample $K \sim \mathrm{Unif}(G)$ and S from its original conditional law given K. Independently sample the randomness that drives the null run. At a prefix where the runs agree, draw the policy action and query once and use that same realized value on both sides. If the action continues, draw the two joint responses from the supplied measurable maximal coupling. After the first response disagreement, complete each run from its own conditional kernels. The Ionescu Tulcea construction gives a probability measure with the prescribed real and null marginals because every kernel in this recursion is measurable.

Let H^0 be the completed null transcript and define

$$\mathsf{Hit} = \left\{ K \in \bigcup_{t \le T} \widehat{W}_t(H^0) \right\}.$$

The null run is independent of (K, S). Conditional on all null randomness, H^0 and its windows are fixed while K is uniform on G. Hypothesis 1 gives

$$\mathbb{P}(\mathsf{Hit} \mid H^0) = \frac{|G \cap \bigcup_{t \le T} \widehat{W}_t(H^0)|}{|G|} \le \frac{\sum_{t \le T} |\widehat{W}_t(H^0)|}{|G|} \le \beta.$$

Averaging over H^0 yields $\mathbb{P}(\mathsf{Hit}) \le \beta$.

Let τ be the first time at which the joint responses disagree, with $\tau = \infty$ if they agree through the horizon. Fix a time t. On $\{\tau \ge t\}$ the two runs share the same history H_t^0 before the response and query q_t. On Hit^c we also have $K \notin W_t(H_t^0, q_t)$. The defining property of maximal coupling therefore gives, conditional on K, S, H_t^0, q_t,

$$\mathbb{P}(\tau = t \mid K, S, H_t^0, q_t, \tau \ge t) = d_{\mathrm{TV}}\big(P_{K,S,t}(\cdot \mid H_t^0, q_t), P_{0,t}(\cdot \mid H_t^0, q_t)\big).$$

Integrate first over S under its original conditional law given K. This is valid because $\mathbf{1}\{\tau \ge t\} \le 1$, so the earlier agreement indicator can be dropped before applying hypothesis 2. We obtain

$$\mathbb{P}(\tau = t, \mathsf{Hit}^c \mid H_t^0, q_t) \le \widehat{\delta}_t(H^0).$$

A sum over t, followed by the pathwise bound in hypothesis 3 and an average over the null run, gives

$$\mathbb{P}(\tau < \infty, \mathsf{Hit}^c) \le \mathbb{E} \sum_{t \le T} \widehat{\delta}_t(H^0) \le \Delta.$$

Since transcript disagreement implies either a hit or a first response mismatch outside all hit windows,

$$\mathbb{P}(\mathsf{Diff}) \le \mathbb{P}(\mathsf{Hit}) + \mathbb{P}(\tau < \infty, \mathsf{Hit}^c) \le \beta + \Delta.$$

On Diff^c the two output kernels receive the same terminal transcript and can use the same output randomness, so their outputs coincide. The coupling inequality gives

$$\mathbb{E}_{K,S} d_{\mathrm{TV}}(Q_{K,S}, Q_0) \le \beta + \Delta.$$

Hypothesis 4 and Tonelli's theorem give

$$\mathbb{E}_{K,S} Q_0(A(S)) = \int \mathbb{P}[x \in A(S)] Q_0(\,\mathrm{d}x) \le \eta.$$

Markov's inequality shows that, outside an instance set of probability at most $(\beta + \Delta)/a$, we have $d_{\mathrm{TV}}(Q_{K,S}, Q_0) < a$. Outside another set of probability at most η/b, we have $Q_0(A(S)) < b$. On the intersection of these two events, $\nu_{K,S}(A(S)) = 1$ and hence

$$d_{\mathrm{TV}}(Q_0, \nu_{K,S}) \ge 1 - Q_0(A(S)) > 1 - b.$$

The triangle inequality then yields $d_{\mathrm{TV}}(Q_{K,S}, \nu_{K,S}) > 1 - a - b$. A union bound over the two exceptional instance sets proves the claim. □

The hit bound conditions on the completed null path, which fixes the window union while leaving the hidden index uniform. The response mismatch bound conditions on the current matching prefix, leaving future null responses random. Dropping the earlier agreement indicator allows Lemma 11 to apply under the original codebook law.

For the deterministic response rule in Appendix C, hypothesis 2 follows by coupling the real and null lists on the same codebook probability space. The joint disagreement event is contained in the union of the $|K_{b_t}(k_t)|$ component events. Lemma 11 bounds each by δ_0, and the total entry cap gives $\sum_t \widehat{\delta}_t \le R_\star \delta_0$. For randomized list responses, hypothesis 2 instead requires a bound on the total variation distance between joint response laws, including their dependence.

F Proof of Theorem 2

Proof. Fix p and I and let $c, d_0 > 0$ be chosen below. Given $d \ge d_0$, $\varepsilon \in (0, 1/4)$ and an integer $R_\star \ge 1$, let $G \subset I$ be the realized candidate rate set of Lemma 17, with $|G| \ge c_I d$ and spacing at least γ/d. Draw $\kappa \sim \mathrm{Unif}(G)$, take $n = e^{d\kappa}$ and a codebook S of n independent uniform vectors in V_d, and equip the instance with the response rule in Appendix C at $\delta_0 = (1000 R_\star)^{-1}$. This specifies $\mathcal{D}_{d,R_\star,\varepsilon}$, its target and its conditional codebook law before the horizon, state space or algorithm is introduced, and Lemma 10 certifies Definition 1.

Now split exhaustively on $\Lambda_\star \le cd$ and $\Lambda_\star > cd$.

Suppose first that $\Lambda_\star \le cd$. Consider the null run, in which every shell component returns $m_{U,k}(x)$. Let k_t^0 denote the center selected at time t along the null transcript. The set of candidate rates exposed at that step is

$$W_t = G \cap \bigcup_{k \in K_{b_t}(k_t^0)} J(k).$$

Apply Lemma 16 with $L = \lambda_-, U = \lambda_+$. The exact endpoints in (18) give $\log(U/L) \le C_H \Lambda_\star$ for a model constant C_H, so we may take $h_R = \lceil C_H \Lambda_\star \rceil \in \mathbb{N}$, and $\Lambda_\star \ge 1$ yields $h_R \le (C_H + 1)\Lambda_\star$. The same bound makes the finite error budget in (19) at most $C_E \Lambda_\star / d$, so choosing c below the fixed separation margin divided by C_E discharges the hypotheses of Lemmas 14 and 15. A zero query path has an empty window family and is handled directly by the support argument below. Therefore $|W_t| \le C(b_t + h_R) \le C'(b_t + \Lambda_\star)$ and, by the second budget in (9),

$$\sum_t |W_t| \;\le\; C' \sum_t \big(b_t + \Lambda_\star\big) \;\le\; C' cd.$$

A sufficiently small choice of c makes this at most $10^{-3}|G|$. The window budget in Theorem 19 is therefore $\beta = 10^{-3}$.

For the response charge, suppose the hidden rate lies outside W_t. Then every shell in the t-th list has $\lambda(n,k) \notin [\lambda_-, \lambda_+]$, so by Lemma 11 each deterministic component differs from its null value at the fixed null query point with probability at most δ_0 under the original codebook law. Since the response rule is deterministic given S, the joint list differs only if some component differs, so the t-th joint mismatch probability is at most $|K_{b_t}(k_t^0)|\delta_0$. Summing and using the first budget in (9) gives $\Delta \le R_\star \delta_0 = 10^{-3}$.

For support separation, for every fixed x,

$$\mathbb{P}[x \in A(S) \mid \kappa] \le n 2^{-d} \le \exp\big[-(\log 2 - \kappa_U)d + 1\big] \le 10^{-3}$$

for all d large, so $\eta = 10^{-3}$. Theorem 19, with $a = b = 1/20$, gives total variation error at least $1 - a - b = 9/10$ with instance probability at least

$$1 - \frac{10^{-3} + 10^{-3}}{1/20} - \frac{10^{-3}}{1/20} = \frac{47}{50} \;\ge\; \frac{9}{10}.$$

Now suppose that $\Lambda_\star > cd$. A single active continue action would contribute $b_t + \Lambda_\star \ge \Lambda_\star > cd$ to a nonnegative accumulated cost, violating the second budget in (9). Hence active continue actions

have zero measure under both the real and the null prefix laws, and the run may be replaced by a process that stops without making a query and has the same output law. Its random window data have $\beta = \Delta = 0$, retaining only $\eta = 10^{-3}$. Theorem 19 with $a = b = 1/20$ gives the same threshold $9/10$ and instance mass $1 - 0 - \frac{10^{-3}}{1/20} = \frac{49}{50} \geq \frac{9}{10}$. This branch uses the same hard law, target and response rule, all fixed before the algorithm. The zero query conclusion follows from the budget premise.

Together the two branches cover every value of $\Lambda_\star$. □

Remark 20 (Budget bounds on every path and almost everywhere). Theorem 2 imposes (9) on every possible transcript path. The proof only uses the weaker consequence that the budgets hold almost everywhere under the original real and null prefix laws. The pathwise premise strictly implies these almost everywhere bounds. Appendix H formalizes this weaker premise, so the formal statement is at least as strong as the version displayed here.

G Proof of Corollary 4

To apply Theorem 2 at fixed half width, we choose the shell cap from the declared query cap.

Proof of Corollary 4. Let $c_{\mathrm{bud}}, d_{0,\mathrm{bud}}$ be the constants of Theorem 2 at the fixed admissible rate interval, and set $c_{\mathrm{fix}} = \min\{1, c_{\mathrm{bud}}/3\}$, $d_0 = \max\{d_{0,\mathrm{bud}}, 2\}$. Fix $d \geq d_0$, $\varepsilon \in (0, 1/4)$, $b \in \{0, \ldots, d\}$, and a predeclared integer $Q_\star \geq 1$ satisfying (13). Put $R_\star = Q_\star(2b+1)$ and $\Lambda_\star = 1 + \log d + \log(1 + R_\star) + p \log(1/\varepsilon)$.

Logarithmic consistency. From (13) and $c_{\mathrm{fix}} \leq 1$ we get $Q_\star \leq d$, hence $1 + R_\star \leq 1 + d(2d+1) \leq (1+d)^3$ for $d \geq 2$ and therefore

$$\log(1 + R_\star) \leq 2\big(\log d + \log(1+b)\big) + O(1), \qquad \Lambda_\star \leq 3\Lambda_b,$$

so that, again by (13), $Q_\star\big(b + \Lambda_\star\big) \leq 3\, Q_\star\big(b + \Lambda_b\big) \leq 3 c_{\mathrm{fix}} d \leq c_{\mathrm{bud}} d$.

Order of quantifiers. Instantiate Theorem 2 at this $R_\star$. Since $R_\star$ is a function of the predeclared $Q_\star$ and b only, the hard family $\mathcal{D}_{d,b,Q_\star,\varepsilon}$, its parameters and its conditional payload law are chosen once $Q_\star$ is fixed. The horizon, state space, and algorithm are quantified afterward.

Transfer of the fixed width policy. Let $T \leq Q_\star$ and let the algorithm's policy be, at every time and history, almost everywhere supported on stop actions and continue queries of half width exactly b. For the resulting clipped band response process, this same fixed width property holds almost everywhere under both original prefix laws. At time t, every preceding continue record and the current query then has half width b. We retain endpoint clipping as an inequality and never replace $|K_b(k)| \leq 2b+1$ by an equality. Since $t + 1 \leq T \leq Q_\star$,

$$\sum_{s \leq t} \big|K_b(k_s)\big| \leq (t+1)(2b+1) \leq Q_\star(2b+1) = R_\star, \qquad \sum_{s \leq t} \big(b + \Lambda_\star\big) \leq Q_\star\big(b + \Lambda_\star\big) \leq c_{\mathrm{bud}} d.$$

These two conclusions establish the almost everywhere budget bounds in (9) under the original real and null laws using the same probabilistic assumptions. Theorem 2 applied to the already chosen hard law yields total variation error at least $9/10$ on a set of instances of probability at least $9/10$. □

Remark 21 (Endpoint clipping). The derivation never replaces $|K_b(k)| \leq 2b+1$ by an equality, so it is valid for centers within b of 0 or d, where the realized band is shorter. The cap condition in (13) states the regime in which the corollary applies.

H Formal Verification

The Lean 4 development uses mathlib. The automated audit finds no uses of the proof placeholders `sorry` or `admit` and no project `axiom` declarations. Every public declaration depends only on Lean's standard classical axioms: propositional extensionality (`propext`, logically equivalent propositions are equal), quotient soundness (`Quot.sound`, related representatives have equal quotient values), and choice (`Classical.choice`, a witness can be selected from a nonempty

Table 3: Correspondence between the statements of this paper and the formal development. All declarations live in the `Radial` namespace.

Statement here	Lean declaration (namespace `Radial`)	Note
Def. 1	`HasNormalizedPerShellLpAccuracy`	real $p \geq 2$, all response randomness
Thm. 2	`totalRadialBudget`	almost everywhere budget bounds
Cor. 4	`fixedBandBudgetIndexed`	$\forall Q_\star\, \exists \mathcal{D}\, \forall \mathcal{A}$ order
Thm. 5	`abs_shellCenter_sub_bounds_of_lowerCompact`	two sided, lower branch
Thm. 19	`randomWindowBudgetCoupling`	finite uniform index, standard Borel payload
Lem. 6	`positiveOccupancyPMFProbability_le_occupancyScale`	
Lem. 7	`uniformShellCovarianceTrace_le_dimension`	
Lem. 8	`expectedGlobalPosteriorSquaredError_le`	
Lem. 9	`expectedPosteriorSquaredError_le`	
Lem. 10	`adversarialShellOracle_lpAccuracy`	every codebook, every valid radius
Lem. 11	`adversarialShellEstimate_pointConsistency`	original codebook law
Lem. 12	`adjacentCenterIdentity`	
Lem. 14	`relevantRadius_mem_lowerCompact`	
Lem. 15	`relevantRadius_mem_compact_symmetric`	
Lem. 16	`bandOccupancyRegion_hitCount_le_compact`	arbitrary bands, incl. straddling $d/2$
Lem. 17	`realizedRateFinset_card_ge_linear`, `realizedRateGrid_adjacent_spacing`	
Prop. 18	`grid_card_le_constant_mul_scaleCoverageBudget`	sum budget with natural number cost constants
	`gaussianAsBandPackingBarrier`	consequence with a maximum half width
Eq. (1)	`nullPosterior_eq_closedForm`	
Eq. (2)	`shellCenter`	definition
Eq. (3)	`log_occupancyScale`	exact identity
Eq. (18)	`occupancyWindow_width`	

type). Each module passes kernel replay, which rechecks its proof terms with Lean's trusted proof checker, in its own memory capped process.

Table 3 links the core budget, geometry, response, and coupling results to their formal declarations. Corollary 3 follows by the arithmetic proof in the main text. Proposition 13 and the translated window interpretation in Appendix D are paper level arguments built on the formal anchor drift theorem.

The declaration `Radial.totalRadialBudget` assumes the budgets hold almost everywhere under the original real and null prefix laws. Theorem 2 imposes them on every possible transcript path, which implies the formal premise.