PARALLELISM, CRITICAL WINDOWS, AND SEPARATIONS AMONG DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A popular selling point of diffusion large language models (dLLMs) is their capacity for *parallelism*: the ability to generate sequences of text far more efficiently than autoregressive models, which require one forward pass per token. Yet among the many competing paradigms for dLLMs, from masked to uniform to Gaussian diffusion, principled understanding of how these different proposals compare in parallelism remains limited. In this work, we initiate a fine-grained comparison of the capacity for parallelism among these three leading approaches. We prove:

- Uniform and Gaussian diffusion can sample in a number of forward passes which scales with the *dual total correlation* of the underlying distribution, a measure of intrinsic complexity which can be much smaller than the context length. Previously, it was only known how to achieve this using masked diffusion (Chen et al., 2026; Lavenant & Zanella, 2025).
- For a certain family of random empirical measures, we show that $\widetilde{\Theta}(\sqrt{d})$ forward passes are necessary and sufficient to sample using uniform or Gaussian diffusion, yet there exist approximate score oracles for which $\widetilde{\Omega}(d)$ forward passes are needed for masked diffusion. This establishes the first provable separation in parallelism between the three prevailing dLLM paradigms.

Contrary to popular intuition that masked diffusions are harder to parallelize because they must commit to token values, the latter separation instead comes from the fact that the *critical windows* in masked diffusion sampling are asymptotically narrower than those in uniform and Gaussian diffusion sampling.

1 INTRODUCTION

Diffusion large language models (dLLMs) have recently emerged as a powerful alternative to autoregressive large language models (LLMs) for generative modeling over discrete domains (Austin et al., 2021; Lou et al., 2023; Sahoo et al., 2024; Nie et al., 2025). Despite the moniker, diffusion language modeling is not one, but a multitude of competing frameworks (Figure 1) all built around the same guiding principle of denoising: given a corruption process that converts data into noise, learn how to undo it in order to transform fresh noise into fresh samples. LLMs are themselves a special case of this paradigm, where the corruption process is right-to-left erasure, and a central question in the theory and practice of dLLMs is whether there is a better choice of corruption process.

One of the most widely touted selling points of moving beyond right-to-left erasure is that it unlocks the ability to perform *few-step generation*: Whereas LLMs must decode one token at a time, dLLMs based on alternative corruption processes can in principle generate a sequence of length d in $o(d)$ forward passes. The cost of doing so, however, remains poorly understood. It is well known that there is *some* statistical price to pay, as such decoding strategies incur *discretization error* from approximating posterior distributions using only their marginals. But the extent to which this price can be rendered negligible depends on how the sampling algorithm at inference time is tuned.

Various empirical works have suggested that this price also depends heavily on the choice of corruption process itself, arguing that few-step generation is easier to achieve for some frameworks than for others (Sahoo et al., 2025; Lee et al., 2026; Hu et al., 2026). For example, it is commonly claimed that corruption processes involving independent erasures, which give rise to the popular paradigm of *masked diffusion models*, are less conducive to few-step generation. The informal reasoning is

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.

that by default, masked diffusion models have to *commit* to token values in discrete steps of the generation process, meaning that mistakes arising from committing multiple tokens in parallel cannot subsequently be corrected (Schiff et al., 2025; Sahoo et al., 2025; 2026). Other popular choices of corruption process, e.g., involving random re-assignments of token values (*uniform diffusion models* (Lou et al., 2023; Sahoo et al., 2025; DiffusionGemma Team, 2026)) or Brownian motion in a latent space (*Gaussian diffusion models* (Li et al., 2022; Dieleman et al., 2022; Roos et al., 2026; Lee et al., 2026; Hu et al., 2026; Dieleman, 2026)), are claimed not to suffer from this as they can progressively revise their output over the course of sampling (see Figure 1).

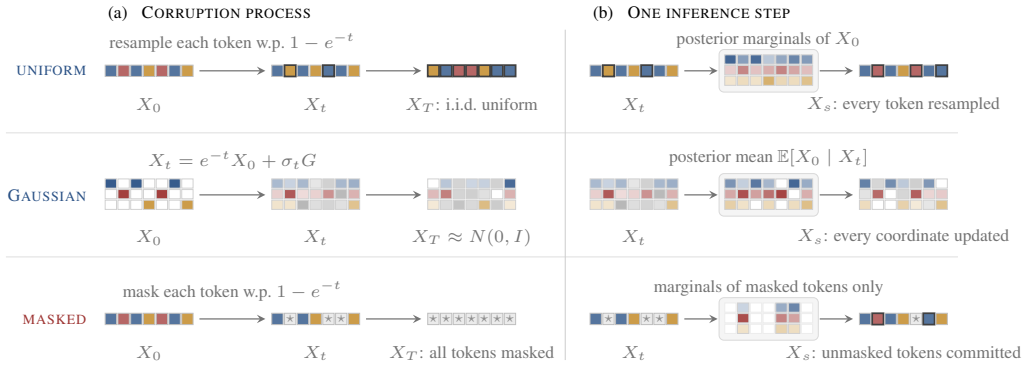


Figure 1: The three dLLM paradigms considered in this work.

Yet these intuitions are largely heuristic, and the foundations of few-step generation under any of these paradigms remain nascent. A flurry of recent theoretical works (Li & Cai, 2026; Chen et al., 2026; Dmitriev et al., 2026a; Wainwright, 2026a; Lavenant & Zanella, 2025; Zhao & Cai, 2026) have suggested that with the right inference schedule, masked diffusion models can achieve few-step generation, where the number of steps scales with an intrinsic complexity measure associated to the data distribution. In contrast, no such results were known for other frameworks like uniform and Gaussian diffusion. In this work, we thus ask:

How does the choice of corruption process affect the few-step generation capabilities of dLLMs?

Framework: score oracle queries. When comparing different dLLM frameworks, one basic figure of merit is the number of forward passes needed to generate an accurate sample. More formally, one forward pass provides the following information. Given a clean sample X_0, the corruption process degrades it into a noisy sample X_t, and one forward pass provides an approximation to the coordinatewise posterior marginals $\mathrm{law}((X_0)_i \mid X_t = z)$ simultaneously for *all coordinates* i, for any noise level t and conditioning z. As these marginals are equivalent, up to affine transformation, to the *annealed score functions* of the data distribution, we will refer to a single forward pass of the model as a query to the *(approximate) score oracle*. We define these oracles formally in Section 2 and Appendix B. When proving *positive results*, we analyze specific algorithms that leverage these oracles; when proving *no-go results*, we rule out *all algorithms* that leverage these oracles.

1.1 Result 1: Scaling with intrinsic complexity

First, we prove that both uniform and Gaussian diffusion can match the scaling established in prior work for masked diffusion. More specifically, we prove the following. Given a categorical distribution q over sequences $X = (X_1, \ldots, X_d)$ in a product space Σ^d, its *dual total correlation* $\mathsf{DTC}(q)$ is, roughly speaking, the total entropy in the distribution which is not explained by the *local* entropies $H(X_i \mid X_{-i})$. Prior work showed that, up to logarithmic factors, masked diffusion models can sample from q in a number of score oracle queries scaling linearly in $\mathsf{DTC}(q)$, which can in general be much smaller than the dimension d.

We begin by showing that such a result is also possible for uniform diffusion.

Theorem 1.1 (Informal, see Theorem 3.1)**.** *Let q be an arbitrary distribution over Σ^d. There is an algorithm that makes $\widetilde{O}(\mathsf{DTC}(q)/\varepsilon)$ queries, where $\widetilde{O}(\cdot)$ hides logarithmic factors in $|\Sigma|$, d, and ε, to an accurate estimate of the* uniform *diffusion score for q and outputs a sample from a distribution $\widehat{q}$ for which $\mathsf{KL}(q \,\|\, \widehat{q}) \leqslant \varepsilon$.*

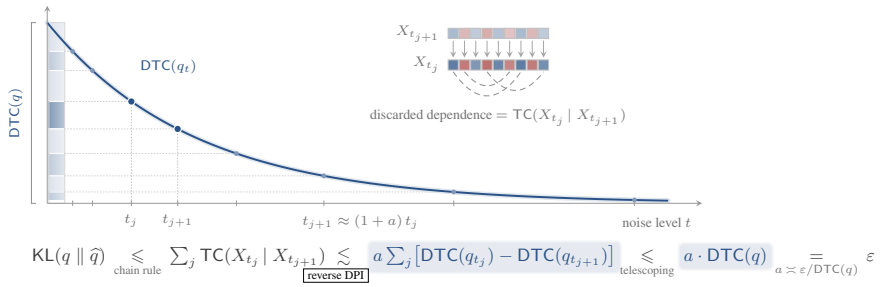


Figure 2: Telescoping argument behind Theorems 1.1 and 1.2. Sampler's KL error is summed over discarded dependencies at each sampling step. Key step is a reverse data-processing inequality bounding each summand, roughly, by step size times drop in $\mathsf{DTC}(q_t)$ in that step.

By a similar mechanism albeit with a more involved argument, we also show:

Theorem 1.2 (Informal, see Theorem 3.2)**.** *There is a sampler with the guarantee of Theorem 1.1 using $\widetilde{O}(\mathsf{DTC}(q)/\varepsilon)$ queries to an accurate estimate of the* Gaussian *diffusion score for q.*

The proof strategy for these results is a marked departure from the one for masked diffusion in Chen et al. (2026). Crucial to that work was an *exact* characterization for the error incurred by the standard algorithm for masked diffusion sampling for any step size schedule; with this, it was enough to substitute an appropriate schedule to obtain the rate $\widetilde{O}(\mathsf{DTC}(q)/\varepsilon)$.

In contrast, for Theorems 1.1 and 1.2, in lieu of an exact characterization, we prove that the error incurred in a single step of the sampler from ignoring conditional correlations can be controlled in terms of the increase in dual total correlation when going from one noise level to a lower noise level. This is proven using a certain *reverse data processing inequality (DPI)* which shows that the distance between two distributions cannot contract too quickly if they are only slightly noised. With this estimate, we can then telescope the DTC increases together and conclude a query complexity that scales linearly in $\mathsf{DTC}(q)$. This argument is illustrated in Figure 2.

1.2 Result 2: Separating masked diffusion from other paradigms

Theorems 1.1 and 1.2 establish that uniform and Gaussian diffusion can adapt to a certain measure of the intrinsic complexity of the data distribution at least as well as masked diffusion. But they leave open whether or not there is a genuine separation among these paradigms.

In our next set of results, we exhibit such a separation by constructing a simple family of *random empirical measures* for which the query complexity of sampling with a masked diffusion score oracle is strictly higher than with a uniform or Gaussian diffusion oracle. More precisely, distributions in this family are given by sampling $e^{\kappa d}$ many points uniformly at random from the Boolean hypercube, where $0 < \kappa < \log 2$ is an (unknown) absolute constant (see Section 4.1 for a formal definition). Recently, Xun & Price (2026) showed that given an approximate Gaussian diffusion score oracle for such a distribution, at least $\widetilde{\Omega}(\sqrt{d})$ queries are necessary even to determine κ to sufficient precision, let alone generate samples close to the distribution. In Appendix I, we prove that this lower bound also applies to uniform diffusion, using similar ideas.

Our main result in this second part of the paper is to show that for these random empirical measures, $\widetilde{O}(\sqrt{d})$ query complexity is actually tight for both uniform and Gaussian diffusion:

Theorem 1.3 (Informal, see Theorems 4.2 and G.1)**.** *For empirical measures q supported on $2^{\Theta(d)}$ random points on the Boolean hypercube, there is an algorithm using $\widetilde{O}(\sqrt{d}/\varepsilon^2)$ queries to any approximate* uniform *diffusion score oracle for q that has score error $1/\mathrm{poly}(d)$ and, with high probability over the support of q, succeeds in sampling from q to total variation error ε. The same holds for any approximate* Gaussian *diffusion score oracle with score error $1/\mathrm{poly}(d)$.*

Notably, for this family, the dual total correlation is $\Theta(d)$, so the scaling achieved in Theorem 1.3 provably goes beyond the dual total correlation scaling implied by Theorems 1.1 and 1.2. In contrast, we show that for the same instance, $\widetilde{\Omega}(d)$ queries are needed for masked diffusion.

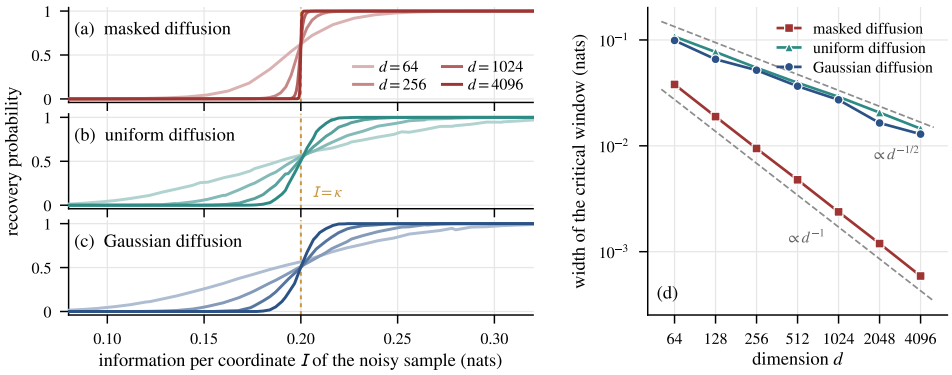


Figure 3: Critical windows for the random empirical measure with $M = e^{\kappa d}$ codewords, $\kappa = 0.2$. (a)–(c) Probability the posterior at a given noise level recovers the planted codeword, plotted against the mutual information I between a uniformly random bit and its corrupted version ($(m/d)\log 2$ for masked diffusion with m revealed coordinates, and $\mathsf{I}_{\mathsf{unif}}(t)$, $\mathsf{I}_{\mathsf{gauss}}(t)$ for uniform and Gaussian diffusion). By Eq. (22), every process has critical level at $\mathsf{I} = \kappa$. (d) Width of transition versus d: decays like $1/d$ for masked diffusion but only like $1/\sqrt{d}$ for uniform and Gaussian diffusion.

Theorem 1.4 (Informal, see Theorem 4.3)**.** *Let q be an empirical measure supported on $2^{\Theta(d)}$ random points on the Boolean hypercube. There is an approximate* masked *diffusion oracle such that any algorithm for sampling making $\widetilde{o}(d)$ queries to this oracle produces samples from a distribution $\widehat{q}$ such that with high probability over the randomness of q, $\mathsf{TV}(q, \widehat{q}) \geq 0.99$.*

These results establish the first provable separation in few-step generation ability among the three main paradigms for diffusion language modeling. We highlight that this manifests through a mechanism fundamentally different from the "commitment" issue mentioned previously. Instead, it comes from a difference in the sharpness of a phase transition arising in the sampling dynamics for masked diffusion versus uniform and Gaussian diffusion. To see this, we note that in both the $\widetilde{\Omega}(\sqrt{d})$ lower bound for uniform and Gaussian diffusion and the $\widetilde{\Omega}(d)$ lower bound for masked diffusion, the key idea is that for random empirical measures, unless one queries the score oracle near the right noise level, the approximation error in the score oracle can be designed to suppress any useful signal about the underlying distribution q. Crucially, however, the range of acceptable noise levels for which this is the case is far narrower for masked diffusion than for uniform and Gaussian diffusion: $O(\log(d)/d)$ fraction of the full noise spectrum for masked diffusion (Proposition H.6) vs. $O(\sqrt{\log(d)/d})$ fraction for the other paradigms (Propositions F.4 and G.4); see Figure 3. As a result $\widetilde{\Omega}(d)$ queries are needed in the former case, whereas only $\widetilde{O}(\sqrt{d})$ suffice in the latter case.

These *critical windows* were previously studied in the context of *speciation transitions* in diffusions (Biroli et al., 2024; Li & Chen, 2024; Li et al., 2025; Sclocchi et al., 2025; Raya & Ambrogioni, 2023). To our knowledge this is the first time they have been used to derive new query complexity bounds for diffusion sampling. Roughly, after identifying the location of the window using $\widetilde{O}(\sqrt{d})$ queries to a uniform or Gaussian diffusion score oracle, we argue that at noise levels above this critical window, the score is well-approximated by that of the uniform distribution, and at noise levels below the critical window, the posterior distribution over the clean sample is so pronounced that one can easily "round" to the correct point in the support. Thus, it suffices to design an algorithm that can efficiently locate and "cross" the window.

2 Preliminaries

2.1 Uniform diffusion

Given alphabet size $S \geq 2$, define $\Sigma = [S]$ and let $\mathcal{X} = \Sigma^d$. Let q be a distribution over $\mathcal{X}$. Let $\Delta(\Sigma)$ denote the probability simplex over Σ. For every $t \geq 0$ define $\alpha_t = e^{-t}$ and $\beta_t = \frac{1-\alpha_t}{S}$, and for $z \in \Sigma$, define the one-coordinate transition kernel

$$K_t^{\mathsf{U}}(\cdot \mid z) = \alpha_t \, \mathbb{1}[z = \cdot] + \beta_t \, .$$

These kernels (K_t^{U}) form a semigroup. Given $X_0 \sim q$, let $(X_t)_{t\geq 0}$ denote the Markov process with transition kernel $(K_t^{\mathsf{U}})^{\otimes d}$, and let $q_t \coloneqq \mathrm{law}(X_t)$.

Given $y \in \mathcal{X}$, $i \in [d]$, and $a \in \Sigma$, let $y^{i\leftarrow a}$ be the point obtained by replacing the i-th coordinate of y with a. The *(uniform diffusion) score at time t* is the function $s_t : \mathcal{X} \to \mathbb{R}_{\geq 0}^{[d]\times\Sigma}$ given by

$$s_t(y)[i,a] = \frac{q_t(y^{i\leftarrow a})}{q_t(y)}\,.$$

We will consider estimated scores $(\widehat{s}_t)$ and quantify their error as follows. First, without loss of generality, we will assume that the estimated scores are *normalized*, i.e., for all t, $y \in \mathcal{X}$, and $i \in [d]$, $\widehat{s}_t(y)[i, y_i] = 1$.

Definition 2.1 (Score error – uniform diffusion)**.** *Given a normalized score estimate $\widehat{s}_t$ at time $t > 0$, its* score error *is $\varepsilon_{\mathsf{unif}}(t) \coloneqq \mathbb{E}_{Y\sim q_t} \sum_{i=1}^{d} \sum_{a\in\Sigma} \psi\big(s_t(Y)[i,a], \widehat{s}_t(Y)[i,a]\big)$. Here, ψ denotes the* (entropic) Bregman divergence *given by $\psi(c,s) \coloneqq s - c + c\log(c/s)$. This is equivalent to the standard score-entropy loss used to train uniform diffusion models (Lou et al., 2023).*

We will assume that $\varepsilon_{\mathsf{unif}}(t) < \infty$ for all $t > 0$, so that $\widehat{s}_t(y)[i,a] > 0$ for all i, a, t, y.

2.2 Gaussian diffusion

We begin by describing Gaussian diffusion for general distributions over $\mathbb{R}^d$, before specializing to embeddings of *discrete distributions*. Let q be a distribution on $\mathbb{R}^d$. Define the *OU process* $(X_t)_{t\geq 0}$ by $\mathrm{d}X_t = -X_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$ with $X_0 \sim q$, where $(B_t)_{t\geq 0}$ is a standard Brownian motion, and let $q_t \coloneqq \mathrm{law}(X_t)$. Defining $\sigma_t = \sqrt{1 - e^{-2t}}$, the transition kernel for this process is given by

$$K_t^{\mathsf{G}}(x \mid z) = \frac{1}{(2\pi\sigma_t^2)^{d/2}} \exp\left(-\frac{\|x - e^{-t}z\|^2}{2\sigma_t^2}\right)\,.$$

These kernels (K_t^{G}) form a semigroup.

The *(Gaussian diffusion) score at time t* is the function $s_t : \mathbb{R}^d \to \mathbb{R}^d$ given by $s_t(x) = \nabla \log q_t(x)$.

We will consider estimated scores $(\widehat{s}_t)$ and quantify their error as follows.

Definition 2.2 (Score error – Gaussian diffusion)**.** *Given score estimate $\widehat{s}_t$ at time t, its L^2* score error *is defined by $\varepsilon_{\mathsf{gauss}}(t) \coloneqq \sigma_t^2 \left(\mathbb{E}_{X_t\sim q_t} \|s_t(X_t) - \widehat{s}_t(X_t)\|_2^2\right)^{1/2}$.*

In this work, we will study Gaussian diffusion in the context of sampling from *discrete distributions*. For general finite alphabet Σ, let q_{pre} be any distribution over $\mathcal{X} = \Sigma^d$. We consider the following standard embedding. Given a sample $A = (A_1, \ldots, A_d)$ from q_{pre} over $\mathcal{X} = [S]^d$, apply the *one-hot encoding* to obtain $X_0 = (X_0^{(1)}, \ldots, X_0^{(d)}) \in \mathbb{R}^{Sd}$, where $X_0^{(i)} = e_{A_i}$ and $e_a \in \mathbb{R}^S$ denotes the a-th standard basis vector. We then take q to denote the pushforward of the data distribution under this one-hot encoding map. We will also consider the special case of binary alphabet $\Sigma = \{-1, 1\}$, in which case we will more directly regard q_{pre} as a distribution over $\mathbb{R}^d$ in the natural way by regarding $\mathcal{X} = \{-1, 1\}^d$ as a subset of $\mathbb{R}^d$.

2.3 Masked diffusion

Given alphabet size $S \geq 2$, unlike in the previous sections we distinguish between $[S]$ and Σ, defining $\Sigma = [S] \cup \{\star\}$, where $\star \notin [S]$ is a special mask symbol. Let $\mathcal{X} = [S]^d$ and $\overline{\mathcal{X}} = \Sigma^d$. Let q be a distribution over $\mathcal{X}$. As in the case of uniform diffusion, let $\alpha_t = e^{-t}$. For $z \in [S]$, define the one-coordinate transition kernel

$$K_t^{\mathsf{M}}(\cdot \mid z) = \alpha_t\, \mathbb{1}[z = \cdot] + (1 - \alpha_t)\, \mathbb{1}[\star = \cdot]\,.$$

These kernels form a semigroup. Given $X_0 \sim q$, let $(X_t)_{t\geq 0}$ denote the Markov process with transition kernel $\left(K_t^{\mathsf{M}}\right)^{\otimes d}$, and let $q_t \coloneqq \mathrm{law}(X_t)$.

For $y \in \overline{\mathcal{X}}$, define the set of revealed coordinates by $I(y) \coloneqq \{i \in [d] : y_i \neq \star\}$. We identify y with the *partial assignment* $(I(y), y_{I(y)})$; conversely, given $I \subseteq [d]$ and $x \in \mathcal{X}$, we write $x^{(I)} \in \overline{\mathcal{X}}$ for the

string obtained from x by masking the coordinates outside I, so that $X_t = X_0^{(I(X_t))}$ and observing X_t is equivalent to observing $(I(X_t), X_{0,I(X_t)})$. We call a partial assignment (I, x_I) *consistent* if $\mathbb{P}[X_{0,I} = x_I] > 0$. Given a consistent partial assignment and an unrevealed coordinate $i \notin I$, define the posterior marginal $q_{i|I}(a \mid x_I) := \mathbb{P}[X_{0,i} = a \mid X_{0,I} = x_I]$ for $a \in [S]$.

The masked diffusion score, formally defined in Appendix B, is equivalent to posterior marginals up to a deterministic time-dependent factor (Lemma B.4), so querying a score estimate at (t, y) is the same as querying an *approximate masked diffusion oracle* $\mathcal{O}$ with the partial assignment $(I, x_I) = (I(y), y_{I(y)})$ and receiving the distributions $\widehat{q}^{\mathcal{O}}_{i|I}(\cdot \mid x_I)$ over $[S]$ for all $i \notin I$; we use both descriptions interchangeably, writing $\widehat{q}_{i|I}$ for $\widehat{q}^{\mathcal{O}}_{i|I}$ when $\mathcal{O}$ is clear from context. On inconsistent partial assignments the posterior marginals are undefined, and the oracle may return arbitrary distributions. The level m is the natural notion of noise level for masked diffusion: unlike for uniform and Gaussian diffusion, it is observable from X_t, and it is the quantity that samplers control in practice. We measure the accuracy of $\mathcal{O}$ by its *score error at level* m, $\varepsilon_{\mathsf{mask}}(\mathcal{O}; q, m)$ (Definition B.5 in Appendix B), which by Lemma B.6 equals

$$\varepsilon_{\mathsf{mask}}(\mathcal{O}; q, m) = \mathbb{E}_{\substack{I \sim \mathrm{Unif}\left(\binom{[d]}{m}\right) \\ X_0 \sim q}} \left[\frac{1}{d-m} \sum_{i \notin I} \mathsf{KL}(q_{i|I}(\cdot \mid X_{0,I}) \,\|\, \widehat{q}^{\mathcal{O}}_{i|I}(\cdot \mid X_{0,I})) \right].$$

2.4 Information-theoretic quantities

Here we formally define the relevant information-theoretic quantities that we use to quantify the intrinsic complexity of a data distribution.

Definition 2.3. *Given random vector $X = (X_1, \ldots, X_d)$, the* total correlation *and* dual total correlation *of X are given by*

$$\mathsf{TC}(X) := \mathsf{KL}(\mathrm{law}(X) \,\|\, \otimes_i \mathrm{law}(X_i)) = \sum_i I(X_i; X_{i+1:d})$$

$$\mathsf{DTC}(X) := H(X) - \sum_i H(X_i \mid X_{-i}).$$

Denote the TC and DTC of X conditioned on *Y by $\mathsf{TC}(X \mid Y)$ and $\mathsf{DTC}(X \mid Y)$.*

The following lemma (proof in Appendix B) bounds the extent to which a coordinatewise channel decreases dual total correlation in terms of a sum of conditional mutual information terms.

Lemma 2.4. *Let $A = (A_1, \ldots, A_d)$ and let $B = (B_1, \ldots, B_d)$ be obtained by applying independent coordinate channels to A. Then*

$$\mathsf{DTC}(A) - \mathsf{DTC}(B) = \mathsf{DTC}(A \mid B) + \sum_{i=1}^{d} I(B_i; A_{-i} \mid B_{-i}).$$

In particular, $\sum_{i=1}^{d} I(B_i; A_{-i} \mid B_{-i}) \leqslant \mathsf{DTC}(A) - \mathsf{DTC}(B)$.

3 DTC-adaptive sampling with uniform and Gaussian diffusion

We now present the formal versions of Theorems 1.1 and 1.2 and outline the argument. Complete proofs are in Appendices C and D, where the theorems are restated as Theorems C.1 and D.1.

Theorem 3.1. *Let q be any distribution on $[S]^d$, and suppose that score estimates $(\widehat{s}_t)$ satisfying $\varepsilon_{\mathsf{unif}}(t) \leqslant \varepsilon_{\mathsf{unif}}$ for all t, along with a number $\overline{\mathsf{DTC}} \geq \mathsf{DTC}(q)$, are available. For every $0 < \varepsilon < 1$, there is a sampler whose output law $\widehat{q}$ satisfies*

$$\mathsf{KL}(q \,\|\, \widehat{q}) \leqslant \varepsilon + O(\varepsilon_{\mathsf{unif}} \cdot \log(dS/\varepsilon)),$$

and which uses $O\left(\left(1 + \frac{\overline{\mathsf{DTC}}}{\varepsilon}\right) \log \frac{edS}{\varepsilon}\right)$ score oracle queries.

We now state the Gaussian diffusion analogue of Theorem 3.1. Recall from Section 2.2 that we consider the pushforward q of a discrete distribution q_{pre} on $\mathcal{X} = \Sigma^d$ to $(\mathbb{R}^S)^d$ via the one-hot encoding. When we refer to the dual total correlation of q, we treat each block of $S = |\Sigma|$ entries as a single coordinate in the definition of $\mathsf{DTC}(q)$, so that $\mathsf{DTC}(q) = \mathsf{DTC}(q_{\mathsf{pre}})$.

Theorem 3.2. *Let q be any one-hot encoding of a distribution q_{pre} on $\mathcal{X}$, and suppose that score estimates $(\widehat{s}_t)$ satisfying $\varepsilon_{\mathsf{gauss}}(t) \leqslant \varepsilon_{\mathsf{gauss}}$ for all t, along with a number $\overline{\mathsf{DTC}} \geq \mathsf{DTC}(q)$, are available. For every $0 < \varepsilon < 1$, there is a sampler whose output law $\widehat{q}$ on $[S]^d$ satisfies*

$$\mathsf{KL}(q_{\mathsf{pre}} \,\|\, \widehat{q}) \leqslant \varepsilon + O(\varepsilon_{\mathsf{gauss}}^2 \cdot \log(dS/\varepsilon)) \,,$$

and which uses $O\Big(\Big(1 + \frac{\overline{\mathsf{DTC}}}{\varepsilon}\Big) \log \frac{edS}{\varepsilon}\Big)$ score oracle queries.

The proof strategy is as follows. Both samplers start at a high noise level and move towards the target by replacing each reverse transition with the product of its coordinatewise conditional marginals, which can be obtained from a score query. With exact scores, the KL cost of this replacement at a step from t to $s < t$ is $\mathsf{TC}(X_s \mid X_t)$. The main lemmas (Lemma 3.3 below for uniform diffusion and Lemma D.2 for Gaussian) bound this cost by $O(t-s)$ times $\mathsf{DTC}(q_s) - \mathsf{DTC}(q_t)$. For instance, in the case of uniform diffusion, we have the following estimate:

Lemma 3.3. *For $0 < s < t$, let $h = t - s$. Then for uniform diffusion,*

$$\mathsf{TC}(X_s \mid X_t) \leqslant \frac{e^{2h} - 1}{1 - e^{-s}} (\mathsf{DTC}(q_s) - \mathsf{DTC}(q_t)) \,.$$

The key ingredient in the proof of this estimate is a reverse data-processing inequality (DPI) for a single token, motivated by the observation (Proposition C.3) that

$$\mathbb{P}[(X_s)_i = a \mid (X_t)_{-i} = y_{-i}] \geq \beta_s \,. \tag{1}$$

Lemma 3.4 (Reverse DPI for uniform diffusion)**.** *Let p, p' be distributions over Σ such that $\min_{a \in \Sigma} \min\{p_a, p'_a\} \geq c/S$ for some $c > 0$. Then $\mathsf{KL}(pK_t^{\mathsf{U}} \,\|\, p'K_t^{\mathsf{U}}) \geq \mathsf{KL}(p \,\|\, p') \cdot \frac{e^{-2t}c}{1-e^{-t}+e^{-t}c}$.*

We apply this to the conditional laws of $(X_s)_i$ given $((X_s)_{i+1:d}, (X_t)_{-i})$ and given $(X_t)_{-i}$ alone, with $c = 1 - e^{-s}$. Combined with the fact that, conditioned on $(X_t)_{-i}$, $(X_s)_{i+1:d} - (X_s)_i - (X_t)_i$ is a Markov chain, Lemma 3.4 yields the inequality shown in Figure 4 (Corollary C.5):

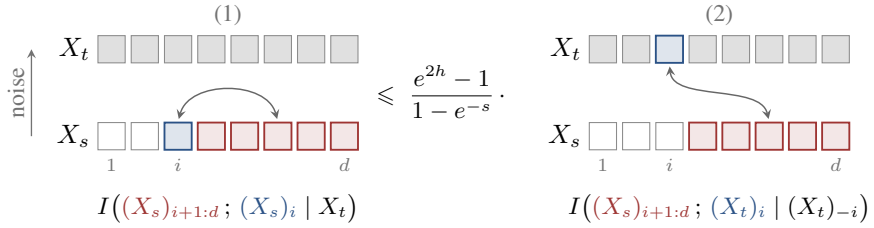


Figure 4: The two conditional mutual informations compared in Corollary C.5. Gray cells are conditioned on, and the arc joins the two arguments of the mutual information.

Summing the LHS in Figure 4 over i yields $\mathsf{TC}(X_s \mid X_t)$, while by Lemma 2.4 the sum of the RHS over i is at most $\mathsf{DTC}(q_s) - \mathsf{DTC}(q_t)$.

For Gaussian diffusion there is no analogue of Eq. (1). Nevertheless, by leveraging a standard identity relating the rate of change of KL divergence along the heat flow to Fisher information, we are able to obtain a reverse DPI for Gaussian diffusion and thus an analogue of Lemma 3.3.

We choose a grid of time-steps (see Figures 5 and 6) for which the pre-factor in front of the DTC decrement in Lemma 3.3, and its Gaussian analogue, is $O(\varepsilon/(1 + \overline{\mathsf{DTC}}))$. In the expression for the overall KL error as a sum of conditional total correlations, by bounding the summands in terms of DTC decrements, we find that the DTC differences telescope, giving $O(\varepsilon)$ discretization error.

4 Separating Masked Diffusion from Other Paradigms

4.1 Construction of Random Empirical Measure

We first describe the lower bound instance, inspired by Xun & Price (2026), deferring details to Appendix E. Fix constants $0 < \kappa_- < \kappa_+ < \log 2$. In the sequel, we will specialize to the case of

binary alphabet, so that $\Sigma = \{\pm 1\}$ and $\mathcal{X} = \{\pm 1\}^d$, and $\beta_t = \frac{1-e^{-t}}{2}$. Let $\mathcal{C} \subseteq \mathcal{X}$ denote the subset of the hypercube defining the codebook, and let $M = |\mathcal{C}|$ denote the size of the codebook.

Throughout, we will take $M = \lceil e^{\kappa d} \rceil$ for constant $\kappa \in [\kappa_-, \kappa_+]$. The compact interval $[\kappa_-, \kappa_+]$ ensures that all the critical values below remain in a fixed compact set.

Definition 4.1. *Given $\mathcal{C} \subseteq \{\pm 1\}^d$, denote the uniform distribution over $\mathcal{C}$ by $q^{\mathcal{C}} := \mathrm{Unif}(\mathcal{C})$. The* random empirical measure *we consider is $q^{\mathcal{C}}$ for $\mathcal{C}$ a uniformly random M-element subset of $\{\pm 1\}^d$. We will often refer to the elements of $\mathcal{C}$ as* codewords.

In all of our bounds in the sequel, we assume that the sampler knows the interval $[\kappa_-, \kappa_+]$ but does not know κ or the specific realization of $\mathcal{C}$.

We define critical times $t^\star_{\mathsf{unif}}$ and $t^\star_{\mathsf{gauss}}$ to be the noise levels at which the per-coordinate mutual information of the uniform, respectively Gaussian, channel equals $d^{-1} \log M$. We provide a formal definition, as well as a specification of a certain interval I_0 containing the critical time, in Appendix E (Eqs. (22) and (23)); our algorithms only query times in I_0.

4.2 $\widetilde{O}(\sqrt{d})$ UPPER BOUNDS FOR UNIFORM AND GAUSSIAN DIFFUSION

Theorem 4.2 (Uniform diffusion with approximate scores)**.** *There are constants $B, C, c_0, d_0 > 0$, depending only on $[\kappa_-, \kappa_+]$, such that the following holds. Let $d \geq d_0$, let $\varepsilon, \delta \in (0, 1/4)$, and set $\Lambda = \log \frac{Cd}{\varepsilon\delta}$. Suppose $\Lambda \leqslant c_0 d$. Given query access to any approximate uniform diffusion score oracle $(\widehat{s}_t)$ for $q^{\mathcal{C}}$ with score error satisfying $\varepsilon_{\mathsf{unif}}(t) \leqslant e^{-B\Lambda} = \left(\frac{\varepsilon\delta}{Cd}\right)^B$ for all $t \in I_0$, there is an algorithm using $O(\sqrt{d\Lambda}/\varepsilon^2)$ oracle queries such that with probability at least $1-\delta$ over $\mathcal{C}$ sampled uniformly from M-element subsets of $\mathcal{X}$, the output law of the algorithm $\widehat{q}^{\mathcal{C}}$ satisfies $\mathsf{TV}(\widehat{q}^{\mathcal{C}}, q^{\mathcal{C}}) \leqslant \varepsilon$.*

In particular, for constant ε and any polynomially small $\delta = d^{-\Theta(1)}$, polynomially small score accuracy suffices, and the algorithm uses $O(\sqrt{d \log d})$ score queries.

Theorem 4.2 is proved in Appendix F. We also prove a Gaussian counterpart (Theorem G.1). To show these, we establish that for both diffusion paradigms, the critical window has width $w \asymp \sqrt{\Lambda/d}$. Above the window, $q^{\mathcal{C}}_t$ is close to ν_t, the noised uniform distribution; below it, X_t identifies the codeword from which it was generated with high probability. Since the codebook size, and hence the critical time, is unknown, the sampler first locates the window; this is the most involved step.

To locate the critical window, it scans the interval I_0 from high to low noise at spacing of order w. To detect when a candidate level lies within the critical window, it uses the following test statistic:

$$T_t(x) = \langle x, m_t(x) - m^0_t(x) \rangle \,,$$

where $m_t(x) = \mathbb{E}[X_0 \mid X_t = x]$ is the denoiser with respect to data distribution $q^{\mathcal{C}}$, and $m^0_t(x)$ is the same w.r.t. $\mathrm{Unif}(\mathcal{X})$. When the noise level is larger (resp. smaller) than the critical window, T_t evaluated on random points from $\mathrm{Unif}(\mathcal{X})$ is typically close to zero (resp. bounded above away from zero). So by drawing $O(\Lambda)$ uniform strings at each t in a grid of $O(1/w)$ noise levels and stopping when there are sufficiently negative values for T_t, the algorithm can identify the critical level to within $O(w)$. The sampler then initializes approximately from ν_t above the window, discretizes the reverse process across it using $O(dw/\varepsilon^2)$ score queries, and recovers a codeword below it.

4.3 $\widetilde{\Omega}(d)$ LOWER BOUND FOR MASKED DIFFUSION

We now consider the same distribution under masked diffusion and exhibit an approximate score oracle for which any sampler requires $\widetilde{\Omega}(d)$ queries, in contrast to uniform and Gaussian diffusion.

Let $\mathcal{K}^{\mathsf{M}}_d := \{k \in \mathbb{Z} : \kappa_- \leqslant k \log 2/d \leqslant \kappa_+\}$, and consider the prior under which $k \sim \mathrm{Unif}(\mathcal{K}^{\mathsf{M}}_d)$ and, given k, $\mathcal{C}$ is a uniformly random 2^k-element subset of $\mathcal{X}$. For a partial assignment (query) (I, x_I), write $\mathsf{match}_I(x_I) := \{y \in \mathcal{C} : y_I = x_I\}$ for the set of codewords matching it. Then (I, x_I) is consistent with $q^{\mathcal{C}}$ if and only if $\mathsf{match}_I(x_I) \neq \varnothing$, in which case

$$q^{\mathcal{C}}_{i|I}(a \mid x_I) = \frac{|\{y \in \mathsf{match}_I(x_I) : y_i = a\}|}{|\mathsf{match}_I(x_I)|}\,, \qquad i \notin I\,.$$

We measure the accuracy of an oracle $\mathcal{O}$ by its level-m score error $\varepsilon_{\mathsf{mask}}(\mathcal{O}; q^{\mathcal{C}}, m)$ from Definition B.5, in the form given by Lemma B.6. This constrains $\mathcal{O}$ only on consistent partial assignments; on inconsistent partial assignments, the oracle can return arbitrary values.

Theorem 4.3 (Masked diffusion lower bound)**.** *Fix $\rho \in (0, 1/4)$. There are constants $c, C, d_0 > 0$, depending only on $[\kappa_-, \kappa_+]$ and ρ, such that the following holds for $d \geq d_0$. Given $\eta \in (0, 1/10)$, let $\ell = \left\lceil \log_2 \frac{Cd}{\eta\rho^2} \right\rceil$. There is a family of oracles $\{\mathcal{O}_{k,\mathcal{C}}\}_{k \in \mathcal{K}_d^{\mathsf{M}}, |\mathcal{C}|=2^k}$ satisfying $\varepsilon_{\mathsf{mask}}(\mathcal{O}_{k,\mathcal{C}}; q^{\mathcal{C}}, m) \leqslant \eta$ for all $0 \leqslant m < d$, $k \in \mathcal{K}_d^{\mathsf{M}}$, and $|\mathcal{C}| = 2^k$, such that every adaptive randomized algorithm $\mathcal{A}$ making at most $c\rho^2 \frac{d}{\ell}$ queries, and which does not know $k, \mathcal{C}$, has output law $\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}$ satisfying*

$$\mathop{\mathbb{P}}_{k,\mathcal{C}}[\mathsf{TV}(\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}, q^{\mathcal{C}}) \geq 1 - \rho] \geq 1 - \rho\,,$$

where $k \sim \mathrm{Unif}(\mathcal{K}_d^{\mathsf{M}})$ and $\mathcal{C} \sim \mathrm{Unif}\big(\binom{\{\pm 1\}^d}{2^k}\big)$.

In particular, for constant ρ and any polynomially small $\eta = d^{-\Theta(1)}$, sampling with at most constant error for at least a constant fraction of the codebooks requires $\Omega(d/\log d)$ queries.

We defer the proof to Appendix H. The argument proceeds as follows. When the number of revealed coordinates m is at most $k - \ell$, for most $\mathcal{C}$, a revealed assignment matches many codewords, and the distribution over any unrevealed coordinate is nearly unbiased so that the true score is close to trivial. When $m > k + \ell$, a random string chosen without knowledge of $\mathcal{C}$ is unlikely to match any codeword, and over such strings the oracle can behave uninformatively as well. These two facts allow us to construct a valid oracle which only leaks information for m in the critical window, and a simple coupling argument between an algorithm that uses this oracle versus an entirely uninformative oracle yields the query complexity lower bound. Crucially, the critical window occupies only an $O(\log d/d)$ fraction of the range of possible values of m, whereas the critical window for uniform and Gaussian diffusion occupies an $O(\sqrt{\log d/d})$ fraction of possible noise levels (Proposition H.6 versus Propositions F.4 and G.4); this is the source of the separation.

5 Outlook

In this work, we compared the parallel sampling capabilities of three leading paradigms for diffusion language modeling. We first showed that the query complexity of uniform and Gaussian diffusion can adapt to the intrinsic complexity of the data distribution quantified by its dual total correlation, matching what was known for masked diffusion (Chen et al., 2026; Lavenant & Zanella, 2025).

On top of establishing parity in this regard, we exhibited a simple family of distributions over which uniform and Gaussian diffusion provably outperform masked diffusion. In contrast with prior heuristic reasoning which suggested that masked diffusion is less amenable to parallelism because it cannot revise the tokens it commits to during sampling, we showed that the origin of this separation comes from a different mechanism. Instead, it is due to the fact that the critical window of noise levels for masked diffusion, over which the posterior transitions from being uninformative to sharply concentrating on a small subset of the support, is asymptotically narrower (by a factor of $\widetilde{\Theta}(\sqrt{d})$) than the analogous windows for uniform and Gaussian diffusion. To our knowledge, this is the first connection between critical windows and discretization error for diffusion sampling.

That said, this work only touches upon one of many important performance axes for diffusion language modeling and should not be interpreted reductively as advocating for one paradigm over another. Indeed, there are various other advantages of masked diffusion, for instance its any-order generation capability, which remain poorly understood in theory but incredibly relevant in practice (Ye et al., 2025; Kim et al., 2025). On the computational side, prior work of Ghio et al. (2024) and Bhatt et al. (2026) compared the algorithmic complexity of denoising under different corruption processes in the context of spin glass-like distributions. Lastly, in the case of Gaussian diffusion, we have only considered the most basic form of embedding, leaving open the exploration of richer encoding maps. Rigorously understanding all of these aspects, and more generally the interplay between corruption process, parallelism, complexity of denoising, and flexibility of inference remains an important open direction, and we leave further exploration of this to future work.

Statement on AI usage

Most of the key ideas of this work were due to the authors from spring and early summer of this year, with assistance from frontier models in a few crucial parts. Most notably, the telescoping argument for the DTC-adaptive rates was suggested by AI. The original proof of the reverse DPI for Gaussian diffusion that was generated by AI was extremely difficult to verify, but further human-AI interaction resulted in the final form of the argument. In an earlier version, we also overlooked the need for the extra Bregman projection step in the uniform diffusion case, which was subsequently suggested by AI. While we had an entirely human-generated argument for our main separation results, AI assisted with simplifying some parts of the argument. It suggested a simpler test statistic for locating the critical window than the one we initially considered. We previously had an alternative argument for sampling below the critical noise level by proving a bound on the relaxed Lipschitz constant of the posterior and then appealing to existing discretization analyses. We believed that this step could be simplified but did not attempt to optimize this part as it did not affect the final result. AI however made the simple but helpful observation that below the critical window, a one-step decoding sufficed. A coding agent also implemented the numerical simulation in Figure 3, with no further guidance beyond the initial prompt seeding it with the experiment of Xun & Price (2026).

In fact, we could use it to prove a considerably more general version of the $\widetilde{O}(\sqrt{d})$ upper bound for Gaussian diffusion than the one that appears here, generalizing from uniform distributions over random subsets of the hypercube to random empirical distributions for any background measure over the unit sphere satisfying concentration of convex Lipschitz functions. We chose not to include this result as Theorem 1.3 already captures the bulk of the intuition.

Frontier models were also used to create all of the figures in this paper, and to catch minor typos and errors in notation and setting of parameters. Beyond these aspects, everything in this paper was human written.

Reproducibility statement

This is a theory paper, and proofs of all mathematical statements are provided in this work. The only code written was for the synthetic experiment giving rise to Figure 3, the full details of which are provided in Appendix K.

References

Beatrice Achilli, Marco Benedetti, Giulio Biroli, and Marc Mézard. Theory of speciation transitions in diffusion models with general class structure. *Journal of Statistical Mechanics: Theory and Experiment*, 2026(4):043304, 2026.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, 2021.

Alankrita Bhatt, Mukur Gupta, Germain Kolossov, and Andrea Montanari. Generating from discrete distributions using diffusions: Insights from random constraint satisfaction problems. *arXiv preprint arXiv:2603.20589*, 2026.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Hongrui Chen and Lexing Ying. Convergence analysis of discrete diffusion model: Exact implementation through uniformization. *arXiv preprint arXiv:2402.08095*, 2024.

Sitan Chen, Kevin Cong, and Jerry Li. Optimal inference schedules for masked diffusion models. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1279–1311. PMLR, 29 Jun–03 Jul 2026. URL https://proceedings.mlr.press/v336/chen26e.html.

Sander Dieleman. Continuous diffusion language models. Blog post, https://sander.ai/2026/08/24/continuous-dlms.html, 2026.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

DiffusionGemma Team. DiffusionGemma technical report. *arXiv preprint arXiv:2608.00146*, 2026.

Daniil Dmitriev, Zhihan Huang, and Yuting Wei. Efficient sampling with discrete diffusion models: Sharp and adaptive guarantees. *arXiv preprint arXiv:2602.15008*, 2026a.

Daniil Dmitriev, Zhihan Huang, and Yuting Wei. Provably adaptive sampling with uniform and remasking discrete diffusion models. *arXiv preprint arXiv:2608.23554*, 2026b.

Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the Sherrington–Kirkpatrick Gibbs measure via algorithmic stochastic localization. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2022.

Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from mean-field Gibbs measures via diffusion processes. *arXiv preprint arXiv:2310.08912*, 2023.

Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The journey, not the destination: How data guides diffusion models. *arXiv preprint arXiv:2312.06205*, 2023.

Davide Ghio, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Sampling with flows, diffusion, and autoregressive neural networks from a spin-glass perspective. *Proceedings of the National Academy of Sciences*, 121(27):e2311810121, 2024.

Te Sun Han and Sergio Verdú. Approximation theory of output statistics. *IEEE Transactions on Information Theory*, 39(3):752–772, 1993.

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

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. ELF: Embedded language flows. *arXiv preprint arXiv:2605.10938*, 2026.

Kelvin Kan, Xingjian Li, Benjamin J Zhang, Tuhin Sahai, Stanley Osher, and Markos A Katsoulakis. Vocabulary-size-independent convergence of discrete diffusion models: Adjoint equations induce the right space. *arXiv preprint arXiv:2605.17232*, 2026.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham M. Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30749–30768. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/kim25ah.html.

Hugo Lavenant and Giacomo Zanella. Error bounds and optimal schedules for masked diffusions with factorized approximations. *arXiv preprint arXiv:2510.25544*, 2025.

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M Boffi, and Jinwoo Kim. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*, 2026.

Gen Li and Changxiao Cai. Breaking ar's sampling bottleneck: Provable acceleration via diffusion language models. *Advances in Neural Information Processing Systems*, 38:11700–11725, 2026.

Marvin Li and Sitan Chen. Critical windows: non-asymptotic theory for feature emergence in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Marvin Li, Aayush Karan, and Sitan Chen. Blink of an eye: a simple theory for feature localization in generative models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35047–35080. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25at.html.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems*, 2022.

Yuchen Liang, Renxiang Huang, Lifeng Lai, Ness Shroff, and Yingbin Liang. Absorb and converge: Provable convergence guarantee for absorbing discrete diffusion models. *arXiv preprint arXiv:2506.02318*, 2025a.

Yuchen Liang, Yingbin Liang, Lifeng Lai, and Ness Shroff. Discrete diffusion models: Novel analysis and new sampler guarantees. In *Advances in Neural Information Processing Systems*, 2025b.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Jianfeng Lu. A mathematical introduction to diffusion models. *arXiv preprint arXiv:2607.01693*, 2026.

Andrea Montanari and Yuchen Wu. Posterior sampling in high dimension via diffusion processes. *arXiv preprint arXiv:2304.11449*, 2023.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Jonathan Niles-Weed and Ilias Zadik. The all-or-nothing phenomenon in sparse tensor pca. *Advances in Neural Information Processing Systems*, 33:17674–17684, 2020.

Le-Tuyet-Nhi Pham, Dario Shariatian, Antonio Ocello, Giovanni Conforti, and Alain Durmus. Bit-level discrete diffusion with Markov probabilistic models: An improved framework with sharp convergence bounds under minimal assumptions. *arXiv preprint arXiv:2502.07939*, 2025.

Le-Tuyet-Nhi Pham, Giovanni Conforti, Zhenjie Ren, and Alain Durmus. Discrete flow matching: Convergence guarantees under minimal assumptions. *arXiv preprint arXiv:2605.08882*, 2026.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. *Advances in Neural Information Processing Systems*, 36:66377–66389, 2023.

Yinuo Ren, Haoxuan Chen, Grant Rotskoff, and Lexing Ying. How discrete and continuous diffusion meet: Comprehensive analysis of discrete diffusion models via a stochastic integral framework. In *International Conference on Learning Representations*, volume 2025, pp. 42904–42941, 2025a.

Yinuo Ren, Haoxuan Chen, Yuchen Zhu, Wei Guo, Yongxin Chen, Grant M Rotskoff, Molei Tao, and Lexing Ying. Fast solvers for discrete diffusion models: Theory and applications of high-order algorithms. In *Advances in Neural Information Processing Systems*, 2025b.

Daan Roos, Oscar Davis, Floor Eijkelboom, Michael Bronstein, Max Welling, Ismail Ilkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. Categorical flow maps. *arXiv preprint arXiv:2602.12233*, 2026.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. In *International Conference on Machine Learning*, 2025.

Subham Sekhar Sahoo, Jean-Marie Lemercier, Zhihan Yang, Justin Deschenaux, Jingyu Liu, John Thickstun, and Ante Jukic. Scaling beyond masked diffusion language models. *arXiv preprint arXiv:2602.15014*, 2026.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. In *International Conference on Learning Representations*, 2025.

Antonio Sclocchi, Alessandro Favero, Noam Itzhak Levi, and Matthieu Wyart. Probing the latent hierarchical structure of data via diffusion models. *arXiv preprint arXiv:2410.13770*, 2024.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1): e2408799121, 2025.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

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

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.

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

Martin J Wainwright. The data geometry of masking diffusion: Certified-optimal schedules via unmasking growth complexity. *arXiv preprint arXiv:2608.13520*, 2026a.

Martin J Wainwright. The information geometry of product-reference discrete diffusion: Interaction growth complexity and optimal scheduling. *arXiv preprint arXiv:2608.28949*, 2026b.

Guanghan Wang, Yair Schiff, Subham Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. *Advances in Neural Information Processing Systems*, 38:147282–147339, 2026.

Andre Wibisono. Mixing time of the proximal sampler in relative fisher information via strong data processing inequality (extended abstract). In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5716–5717. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/wibisono25a.html.

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

Jiacheng Ye, Jiahui Gao, Shansan Gong, Lin Zheng, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Beyond autoregression: Discrete diffusion for complex reasoning and planning. In *International Conference on Learning Representations*, volume 2025, pp. 77875–77898, 2025.

Zikun Zhang, Zixiang Chen, and Quanquan Gu. Convergence of score-based discrete diffusion models: A discrete-time analysis. In *International Conference on Learning Representations*, 2025.

Yunxiao Zhao and Changxiao Cai. Adaptation to intrinsic dependence in diffusion language models. *arXiv preprint arXiv:2602.20126*, 2026.

TABLE OF CONTENTS

A Related work **16**

B Omitted preliminaries **16**

B.1 Uniform diffusion 17

B.2 Gaussian diffusion 17

B.3 Masked diffusion 18

B.4 Information-theoretic quantities 19

C Uniform diffusion can scale with dual total correlation **19**

C.1 Conditional TC versus DTC decrement 20

C.2 Score error and approximate reverse kernel 22

C.3 Sampler and telescoping argument 23

D Gaussian diffusion can scale with dual total correlation **26**

D.1 Conditional TC versus DTC decrement 26

D.2 Reverse DPI for Gaussian diffusion 27

D.3 Score error and approximate reverse kernel 28

D.4 Sampler and telescoping argument 30

E Construction of random empirical measure **32**

F $\widetilde{O}(\sqrt{d})$ upper bound for uniform diffusion **33**

F.1 Quantifying the critical window 34

F.2 Identifying the critical window 36

F.3 Sampling after locating the window 40

G $\widetilde{O}(\sqrt{d})$ upper bound for Gaussian diffusion **42**

G.1 Quantifying the critical window 43

G.2 Identifying the critical window 46

G.3 Sampling after locating the window 50

H $\widetilde{\Omega}(d)$ lower bound for masked diffusion **54**

H.1 Adversarial oracle construction 55

H.2 Distribution indistinguishability: the lower bound proof 56

I $\widetilde{\Omega}(\sqrt{d})$ lower bound for uniform diffusion **58**

I.1 Adversarial oracle construction 60

I.2 Distribution indistinguishability: the lower bound proof 62

J Deferred proofs **64**

J.1 Proofs from Section 2 64
J.2 Proofs from Section C 65
J.3 Proofs from Section D 66
J.4 Proofs from Section F 67
J.5 Proofs from Section G 70

K Details of simulation for Figure 3 **75**

A Related work

Theory for discrete diffusion sampling. Campbell et al. (2022) formulated continuous-time discrete denoising through continuous-time Markov chains. Chen and Ying (2024) used uniformization to implement the reverse chain associated with an estimated score and obtained total-variation and KL guarantees on the hypercube, notably obtaining zero discretization error in $O(d)$ uniform diffusion score oracle queries. Li and Cai (2026) initiated the theoretical study of the query complexity of masked diffusion as a function of information-theoretic properties of the data distribution. Chen, Cong, and Li (2026) and Lavenant and Zanella (2025) derived an exact characterization for the sampling error in terms of the data distribution's *information curve*, which can be used to obtain query complexity bounds scaling linearly in the total correlation and dual total correlation. Closely related to our lower bound for masked diffusion, Chen et al. (2026) also gave an $\Omega(d)$ lower bound with a different construction based on MDS codes. Dmitriev et al. (2026a) and Zhao and Cai (2026) proved sharp convergence guarantees for standard uniform and masking τ-leaping schemes, showing that the latter can adapt to the total and dual total correlation of the distribution.

Separately, Ren et al. (2025a) developed a stochastic-integral framework for KL analysis of time-discretized reverse processes. Building on this framework, Ren et al. (2025b) developed and analyzed higher-order solvers for discrete diffusion. More broadly, there is a large body of orthogonal work that analyzes the convergence of CTMC-based discrete diffusion samplers, with particular attention to the dependence on the alphabet size (Zhang et al., 2025; Pham et al., 2025; Liang et al., 2025a;b; Pham et al., 2026; Kan et al., 2026).

Additionally, there is an extensive theoretical literature on *Gaussian* diffusion models — see the recent notes of Lu (2026) for a comprehensive overview. Most relevant to our work is the lower bound of Xun and Price (2026) showing that $\widetilde{\Omega}(\sqrt{d})$ queries are necessary for Gaussian diffusion. The second half of this paper is based on one of the constructions studied in that work. In addition, there have been a number of works on *algorithmic stochastic localization* for sampling from distributions over the Boolean hypercube (El Alaoui et al., 2022; Montanari & Wu, 2023; El Alaoui et al., 2023), essentially by regarding the domain as a subset of Euclidean space, performing Gaussian diffusion, and rounding at the end.

Critical windows and speciation. Numerous works (Georgiev et al., 2023; Raya & Ambrogioni, 2023; Sclocchi et al., 2024; 2025; Biroli et al., 2024; Li & Chen, 2024; Li et al., 2025; Achilli et al., 2026) have studied an intriguing property of real-world diffusion models whereby key aspects of the final generation emerge over a narrow range of times in the reverse process. These works characterize this behavior as a phase transition, interchangeably termed a *critical window* or a *speciation transition*, and provide various probabilistic models of data under which this arises. This phenomenon is the same one that manifests in the context of the random empirical measures considered in Xun & Price (2026) and that we study in the second part of this work, which is in turn closely related to concepts from the statistics and information theory literature like the all-or-nothing phenomenon in high-dimensional inference (Niles-Weed & Zadik, 2020) and channel resolvability and soft covering (Han & Verdú, 1993).

Concurrent work. Recently, Dmitriev, Huang, and Wei (2026b) and Wainwright (2026b) also independently established DTC-adaptive bounds for uniform diffusion. Their proof strategies bear similarities to the proof of our Theorem 1.1, in particular bounding the sampling error by a telescoping sum of differences between dual total correlation at successive noise levels. Interestingly, their works give a characterization of the discretization error for *any* product forward process in terms of the change in the mutual information between one coordinate and the remaining coordinates at other times. In contrast to our work, Wainwright (2026b) further leverages this characterization to give a rate adaptive to the total correlation, and Dmitriev et al. (2026b) proves a DTC-adaptive bound for *remasking* diffusion (Wang et al., 2026). In contrast to these two works, in our work we prove a DTC-adaptive bound for *Gaussian* diffusion and also study separations among the various frameworks in terms of score oracle query complexity.

B Omitted preliminaries

This appendix collects the parts of Section 2 that were omitted from the main text.

B.1 Uniform diffusion

The score naturally admits an interpretation both in terms of posterior marginals and in terms of Gibbs sampling marginals:

Lemma B.1. *Given a vector $v \in \mathbb{R}^S_{>0}$ and given $h > 0$ and $b \in \Sigma$, define probability distributions $p^{\mathsf{post}}_{h,b}, p^{\mathsf{gibbs}}_h$ over Σ by*

$$p^{\mathsf{post}}_{h,b}[v](a) := \frac{\alpha_h \mathbb{1}[a = b] + \beta_h}{\alpha_h}\Big(v_a - \beta_h \sum_{a' \in \Sigma} v_{a'}\Big).$$

$$p^{\mathsf{gibbs}}_h[v](a) := \frac{1}{\alpha_h}\Big(\frac{v_a}{\sum_{a' \in \Sigma} v_{a'}} - \beta_h\Big).$$

Let $h = t - s$. Given $y \in \mathcal{X}$,

$$\mathbb{P}[X^i_s = a \mid X_t = y] = p^{\mathsf{post}}_{h,y_i}[s_t(y)[i, \cdot]](a),$$

and

$$\mathbb{P}[X^i_s = a \mid (X_t)_{-i} = y_{-i}] = p^{\mathsf{gibbs}}_h[s_t(y)[i, \cdot]](a).$$

Here, vectors with the subscript $-i$ denote the vectors obtained by deleting coordinate i.

We defer the proof to Appendix J.

In Section C.2, we discuss a robust analogue of Lemma B.1 for converting an approximate score into a valid estimate for the posterior marginals.

For lower-bound arguments, it will be convenient to make the score estimate and target distribution explicit, using the oracle language. An *approximate uniform diffusion oracle* $\mathcal{O}$, queried at (t, y), returns a score estimate $\widehat{s}^{\mathcal{O}}_t(y)$. We write

$$\varepsilon_{\mathsf{unif}}(\mathcal{O}; q, t) := \mathbb{E}_{Y \sim q_t} \sum_{i=1}^{d} \sum_{a \in \Sigma} \psi\big(s_t(Y)[i, a], \widehat{s}^{\mathcal{O}}_t(Y)[i, a]\big)$$

for the corresponding score error. When q and $\mathcal{O}$ are clear from context, we abbreviate $\varepsilon_{\mathsf{unif}}(\mathcal{O}; q, t)$ by $\varepsilon_{\mathsf{unif}}(t)$ and write $\widehat{s}_t$ for $\widehat{s}^{\mathcal{O}}_t$.

Finally, for the case of binary alphabet $\Sigma = \{-1, +1\}$, the forward likelihood ratio has a convenient overlap form which will be useful in Section F.

Lemma B.2 (Binary likelihood ratio)**.** *For every $x, z \in \{-1, +1\}^d$,*

$$2^d (K^{\mathsf{U}}_t)^{\otimes d}(x \mid z) = \exp\left(\sum_{i=1}^{d} \log(1 + e^{-t} x_i z_i)\right) = (1 + e^{-t})^d \tanh(t/2)^{d_H(x,z)},$$

where $d_H(x, z)$ denotes the Hamming distance between x and z.

Proof. Coordinatewise, $2K^{\mathsf{U}}_t(x_i \mid z_i) = 1 + e^{-t} x_i z_i$. Multiplying these identities gives the first equality. A matching coordinate contributes $1 + e^{-t}$ and a mismatching coordinate contributes $1 - e^{-t} = (1 + e^{-t}) \tanh(t/2)$, which gives the second. □

B.2 Gaussian diffusion

The score naturally admits an interpretation in terms of posterior expectations:

Lemma B.3 (Tweedie's formula)**.** *For any $x \in \mathbb{R}^d$ and $t > 0$,*

$$\mathbb{E}[X_0 \mid X_t = x] = e^t x + e^t \sigma_t^2 s_t(x).$$

By Lemma B.3, $\sigma_t^2\big(s_t(x) - \widehat{s}_t(x)\big) = e^{-t}\big(\mathbb{E}[X_0 \mid X_t = x] - \widehat{m}_t(x)\big)$ for the denoiser $\widehat{m}_t(x) := e^t x + e^t \sigma_t^2 \widehat{s}_t(x)$ associated to $\widehat{s}_t$, so $\varepsilon_{\mathsf{gauss}}(t)^2 = e^{-2t}\, \mathbb{E}_{X_t \sim q_t} \|\widehat{m}_t(X_t) - \mathbb{E}[X_0 \mid X_t]\|_2^2$ is, up to the factor e^{-2t}, the excess risk of $\widehat{m}_t$ for the denoising objective $\mathbb{E}\|\widehat{m}_t(X_t) - X_0\|_2^2$ used to train Gaussian diffusion models (Sohl-Dickstein et al., 2015; Vincent, 2011; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021), whose minimizer is the posterior mean.

B.3 MASKED DIFFUSION

Given $y \in \overline{\mathcal{X}}$, $i \notin I(y)$, and $a \in [S]$, let $y^{i \leftarrow a}$ be given by replacing the i-th coordinate of y with a. The *(masked diffusion) score at time t* is the function $s_t : \overline{\mathcal{X}} \to \mathbb{R}_{\geq 0}^{[d] \times [S]}$ given by

$$s_t(y)[i, a] \coloneqq \frac{q_t(y^{i \leftarrow a})}{q_t(y)}, \qquad i \notin I(y), \quad a \in [S],$$

and $s_t(y)[i, a]$ is undefined for $i \in I(y)$.

The masked diffusion score has a direct interpretation in terms of posterior marginals.

Lemma B.4. *Let $y \in \overline{\mathcal{X}}$ be consistent and let $I = I(y) \subsetneq [d]$. For any $i \notin I$ and $a \in [S]$,*

$$s_t(y)[i, a] = \frac{\alpha_t}{1 - \alpha_t} q_{i|I}(a \mid y_I).$$

Proof. Observe that $q_{i|I}(a \mid y_I) = \frac{\mathbb{P}[X_{0,i} = a, X_{0,I} = y_I]}{\mathbb{P}[X_{0,I} = y_I]}$, while $q_t(y^{i \leftarrow a}) = \alpha_t^{|I|+1}(1 - \alpha_t)^{d-|I|-1} \mathbb{P}[X_{0,i} = a, X_{0,I} = y_I]$ and $q_t(y) = \alpha_t^{|I|}(1 - \alpha_t)^{d-|I|} \mathbb{P}[X_{0,I} = y_I]$. Taking the ratio gives the desired result. □

We will consider estimated scores $(\widehat{s}_t)$ and quantify their error as follows. First, we will assume that the estimated scores are *consistent*, i.e., that for all t, $y \in \overline{\mathcal{X}}$, and $i \notin I(y)$, $\sum_{a \in [S]} \frac{1-\alpha_t}{\alpha_t} \widehat{s}_t(y)[i, a] = 1$, and define the *estimated posterior marginals* by $\widehat{q}_{i|I}(a \mid y_I) \coloneqq \frac{1-\alpha_t}{\alpha_t} \widehat{s}_t(y)[i, a]$, which we assume do not depend on t, as is the case for the true posterior marginals. By Lemma B.4, querying such a score estimate at (t, y) is then the same as querying an *approximate masked diffusion oracle $\mathcal{O}$* with the partial assignment $(I, x_I) = (I(y), y_{I(y)})$ and receiving the distributions $\widehat{q}^{\mathcal{O}}_{i|I}(\cdot \mid x_I)$ over $[S]$ for all $i \notin I$; we use both descriptions interchangeably, writing $\widehat{q}_{i|I}$ for $\widehat{q}^{\mathcal{O}}_{i|I}$ when $\mathcal{O}$ is clear from context. On inconsistent partial assignments the posterior marginals are undefined, and the oracle may return arbitrary distributions.

Definition B.5 (Score error – masked diffusion)**.** *Given consistent score estimate $\widehat{s}_t$ at time t, its* score error *is defined as*

$$\varepsilon_{\mathsf{mask}}(t) \coloneqq \frac{1 - \alpha_t}{\alpha_t} \mathbb{E}_{Y \sim q_t} \left[\frac{1}{d - |I(Y)|} \sum_{i \in [d] \setminus I(Y)} \sum_{a \in [S]} \psi(s_t(Y)[i, a], \widehat{s}_t(Y)[i, a]) \right],$$

with the convention that the bracket is 0 when $I(Y) = [d]$. For the associated oracle $\mathcal{O}$ and $0 \leqslant m < d$, the score error of $\mathcal{O}$ at level m *is, for any $t > 0$, the same quantity conditioned on $|I(Y)| = m$ (it does not depend on t, see below):*

$$\varepsilon_{\mathsf{mask}}(\mathcal{O}; q, m) \coloneqq \frac{1 - \alpha_t}{\alpha_t} \mathbb{E}_{\substack{I \sim \mathrm{Unif}\left(\binom{[d]}{m}\right) \\ X_0 \sim q}} \left[\frac{1}{d - m} \sum_{i \notin I} \sum_{a \in [S]} \psi\Big(s_t\big(X_0^{(I)}\big)[i, a], \widehat{s}_t\big(X_0^{(I)}\big)[i, a]\Big) \right].$$

Since $I(X_t)$ is a uniformly random subset of size $|I(X_t)| \sim \mathrm{Bin}(d, \alpha_t)$ independent of X_0, we have $\varepsilon_{\mathsf{mask}}(t) = \mathbb{E}_{m \sim \mathrm{Bin}(d, \alpha_t)}[\varepsilon_{\mathsf{mask}}(\mathcal{O}; q, m)\, \mathbb{1}[m < d]]$, and $\varepsilon_{\mathsf{mask}}(\mathcal{O}; q, m)$ does not depend on t (see Lemma B.6). The level m is the natural notion of noise level for masked diffusion: unlike for uniform and Gaussian diffusion, it is observable from X_t, and it is the quantity that samplers control in practice.

The following lemma reinterprets the masked diffusion score error in terms of the estimated posterior marginals.

Lemma B.6. *Let $(\widehat{s}_t)$ be consistent, with associated oracle $\mathcal{O}$. Then for every $0 \leqslant m < d$,*

$$\varepsilon_{\mathsf{mask}}(\mathcal{O}; q, m) = \mathbb{E}_{\substack{I \sim \mathrm{Unif}\left(\binom{[d]}{m}\right) \\ X_0 \sim q}} \left[\frac{1}{d - m} \sum_{i \notin I} \mathsf{KL}(q_{i|I}(\cdot \mid X_{0,I}) \,\|\, \widehat{q}^{\mathcal{O}}_{i|I}(\cdot \mid X_{0,I})) \right]. \tag{2}$$

In particular, $\varepsilon_{\mathsf{mask}}(\mathcal{O}; q, m)$ depends on $\mathcal{O}$ only through the marginals it returns on consistent partial assignments with m revealed coordinates.

We defer the proof to Appendix J. The right-hand side of Eq. (2) is, up to an additive term independent of $\mathcal{O}$ and the weighting over levels m, the cross-entropy loss $\sum_{i \notin I} -\log \widehat{q}_{i|I}(X_{0,i} \mid X_{0,I})$ on masked positions, commonly used to train masked diffusion models (Sahoo et al., 2024; Shi et al., 2024).

B.4 Information-theoretic quantities

Here we restate and prove the identity relating sums of conditional mutual informations to the DTC decrement.

Lemma 2.4. *Let $A = (A_1, \dots, A_d)$ and let $B = (B_1, \dots, B_d)$ be obtained by applying independent coordinate channels to A. Then*

$$\mathsf{DTC}(A) - \mathsf{DTC}(B) = \mathsf{DTC}(A \mid B) + \sum_{i=1}^{d} I(B_i; A_{-i} \mid B_{-i})\,.$$

In particular, $\sum_{i=1}^{d} I(B_i; A_{-i} \mid B_{-i}) \leqslant \mathsf{DTC}(A) - \mathsf{DTC}(B)$.

Proof. By definition,

$$\begin{aligned}
&\mathsf{DTC}(A) - \mathsf{DTC}(B) - \mathsf{DTC}(A \mid B) \\
&= H(A) - H(B) - H(A \mid B) - \sum_i \Big(H(A_i \mid A_{-i}) - H(B_i \mid B_{-i}) - H(A_i \mid A_{-i}, B) \Big)\,. \qquad (3)
\end{aligned}$$

Note that

$$H(A) - H(B) - H(A \mid B) = -H(B \mid A) = -\sum_i H(B_i \mid A_i)\,,$$

where the second equality follows because B is obtained through a product channel applied to A. For the same reason,

$$H(A_i \mid A_{-i}, B) = H(A_i \mid A_{-i}, B_i)\,,$$

and

$$\begin{aligned}
&H(A_i \mid A_{-i}) - H(A_i \mid A_{-i}, B_i) + H(B_i \mid A_i) \\
&\qquad = H(A_i \mid A_{-i}) - H(A_i \mid A_{-i}, B_i) + H(B_i \mid A_i, A_{-i}) = H(B_i \mid A_{-i})\,.
\end{aligned}$$

Substituting these into Eq. (3),

$$\mathsf{DTC}(A) - \mathsf{DTC}(B) - \mathsf{DTC}(A \mid B) = \sum_i \Big(H(B_i \mid B_{-i}) - H(B_i \mid A_{-i}) \Big)$$

Finally, note that $H(B_i \mid A_{-i}) = H(B_i \mid A_{-i}, B_{-i})$, so the above can be written as

$$= \sum_i I(B_i; A_{-i} \mid B_{-i})\,,$$

as claimed. □

C Uniform diffusion can scale with dual total correlation

In this section we prove our first main result that uniform diffusion can achieve query complexity scaling with the dual total correlation of the distribution, in analogy to what was previously shown for masked diffusion Chen et al. (2026).

Theorem C.1. *Let q be any distribution on $[S]^d$, and suppose that score estimates $(\widehat{s}_t)$ satisfying $\varepsilon_{\mathsf{unif}}(t) \leqslant \varepsilon_{\mathsf{unif}}$ for all t, along with a number $\overline{\mathsf{DTC}} \geq \mathsf{DTC}(q)$, are available. For every $0 < \varepsilon < 1$, there is a sampler whose output law $\widehat{q}$ satisfies*

$$\mathsf{KL}(q \,\|\, \widehat{q}) \leqslant \varepsilon + O(\varepsilon_{\mathsf{unif}} \cdot \log(dS/\varepsilon))\,,$$

and which uses

$$O\left(\left(1 + \frac{\overline{\mathsf{DTC}}}{\varepsilon} \right) \log \frac{edS}{\varepsilon} \right)$$

score oracle queries.

In Section C.1, we prove the key estimate bounding the conditional TC in terms of the amount by which the DTC decreases when going from a lower noise level to a higher noise level, enabling a telescoping argument. In Section C.2, we control the effect of score error. Finally, in Section C.3, we describe the sampling algorithm and complete the proof of Theorem C.1.

C.1 Conditional TC versus DTC decrement

The main result of this section is an upper bound on the conditional TC, $\mathsf{TC}(X_s \mid X_t)$, in terms of the amount by which the dual total correlation decreases when going from noise level s to a higher noise level t:

Lemma C.2. *For $0 < s < t$, let $h = t - s$. Then*

$$\mathsf{TC}(X_s \mid X_t) \leqslant \frac{e^{2h} - 1}{1 - e^{-s}} (\mathsf{DTC}(q_s) - \mathsf{DTC}(q_t)) \,.$$

We first make a simple but key observation about the Gibbs sampling marginals which will drive the proof of the lemma above:

Proposition C.3. *For $0 < s \leqslant t$ and for all $i \in [d], a \in \Sigma, y \in \mathcal{X}$,*

$$\mathbb{P}[(X_s)_i = a \mid (X_t)_{-i} = y_{-i}] \geq \beta_s \,.$$

Proof. If in addition to $(X_t)_{-i} = y_{-i}$, one also conditions on all of X_0, then $(X_s)_i$ and $(X_t)_{-i}$ become conditionally independent. But conditioned on X_0, the probability that $(X_s)_i = a$ is, by definition of the transition kernel K_s^{U}, lower bounded by β_s. Averaging this bound over the conditional law of X_0 given $(X_t)_{-i} = y_{-i}$, we prove the claim. □

We next establish the following helper lemma, motivated by the observation above. It shows that if two distributions over Σ place some nonzero amount of mass on every element, then applying a small amount of noise does not decrease the KL divergence between them by too much, yielding a reverse data-processing inequality (DPI) for uniform diffusion on a single token.

Lemma C.4 (Reverse DPI for uniform diffusion)**.** *Let p, p' be distributions over Σ such that $\min_{a \in \Sigma} \min\{p_a, p'_a\} \geq c/S$ for some $c > 0$. Then*

$$\mathsf{KL}(pK_t^{\mathsf{U}} \parallel p'K_t^{\mathsf{U}}) \geq \mathsf{KL}(p \parallel p') \cdot \frac{e^{-2t}c}{1 - e^{-t} + e^{-t}c} \,.$$

Proof. Because for any $x, y > 0$, $x \log(x/y) - x + y = (x - y)^2 \int_0^1 \frac{1-\theta}{y + \theta(x-y)} \, \mathrm{d}\theta$, we have

$$\mathsf{KL}(p \parallel p') = \sum_a (p_a - p'_a)^2 \int_0^1 \frac{1 - \theta}{p'_a + \theta(p_a - p'_a)} \, \mathrm{d}\theta \,. \tag{4}$$

By assumption, for all $a \in \Sigma$, $p'_a + \theta(p_a - p'_a) \geq c/S$ for all $\theta \in [0, 1]$, and thus

$$e^{-t}(p'_a + \theta(p_a - p'_a)) + \frac{1 - e^{-t}}{S} \leqslant \left(e^{-t} + \frac{1 - e^{-t}}{c}\right)(p'_a + \theta(p_a - p'_a)) \,.$$

Additionally, $((pK_t^{\mathsf{U}})_a - (p'K_t^{\mathsf{U}})_a)^2 = e^{-2t}(p_a - p'_a)^2$. Applying the same identity that gives rise to Eq. (4), we conclude that

$$\mathsf{KL}(pK_t^{\mathsf{U}} \parallel p'K_t^{\mathsf{U}}) \geq \frac{e^{-2t}}{e^{-t} + (1 - e^{-t})/c} \mathsf{KL}(p \parallel p') \,,$$

as desired. □

We can use the above estimate to compare two quantities: (1) the mutual information between a coordinate i of X_s and some other coordinates of X_s, conditioned on the sequence X_t at a higher noise level t, and (2) the same quantity except $(X_s)_i$ is replaced with $(X_t)_i$ and only the remaining coordinates of X_t are conditioned upon. See Figure 4 for an illustration.

Corollary C.5. *For $0 < s < t$, let $h = t - s$. Then for every $i \in [d-1]$,*

$$I((X_s)_{i+1:d}; (X_s)_i \mid X_t) \leqslant I((X_s)_{i+1:d}; (X_t)_i \mid (X_t)_{-i}) \cdot \frac{e^{2h}-1}{1-e^{-s}}\,. \tag{5}$$

Proof. Conditioned on $(X_t)_{-i}$, we have a Markov chain

$$(X_s)_{i+1:d} - (X_s)_i - (X_t)_i\,,$$

as $(X_t)_i$ is generated from $(X_s)_i$ by applying K_h^{U} with randomness independent of $((X_s)_{i+1:d}, (X_t)_{-i})$. Note that $I(A; B \mid C) = \mathbb{E}_{A,C}\mathsf{KL}(\mathrm{law}(B \mid A, C) \parallel \mathrm{law}(B \mid C))$, so

$$I((X_s)_{i+1:d}; (X_t)_i \mid (X_t)_{-i}) = \mathbb{E}\ \mathsf{KL}(\mathrm{law}((X_t)_i \mid (X_s)_{i+1:d}, (X_t)_{-i}) \parallel \mathrm{law}((X_t)_i \mid (X_t)_{-i})) \tag{6}$$

$$I((X_s)_{i+1:d}; (X_s)_i \mid (X_t)_{-i}) = \mathbb{E}\ \mathsf{KL}(\mathrm{law}((X_s)_i \mid (X_s)_{i+1:d}, (X_t)_{-i}) \parallel \mathrm{law}((X_s)_i \mid (X_t)_{-i}))\,. \tag{7}$$

Conditioned on any realization of $(X_s)_{i+1:d}$ and $(X_t)_{-i}$, the conditional distribution on $(X_s)_i$ is given by the mixture

$$\mathrm{law}((X_s)_i \mid (X_s)_{i+1:d}, (X_t)_{-i}) = e^{-s}\mathrm{law}((X_0)_i \mid (X_s)_{i+1:d}, (X_t)_{-i}) + \frac{1-e^{-s}}{S}\,,$$

so if we take p and p' in Lemma C.4 to be $\mathrm{law}((X_s)_i \mid (X_s)_{i+1:d}, (X_t)_{-i})$ and $\mathrm{law}((X_s)_i \mid (X_t)_{-i})$, and if we take t therein to be $h = t - s$, then by the display above and Proposition C.3, the condition in the lemma is satisfied with $c = 1 - e^{-s}$, and we conclude from this and Eqs. (6) and (7) after averaging over all conditionings that

$$I((X_s)_{i+1:d}; (X_t)_i \mid (X_t)_{-i}) \geq I((X_s)_{i+1:d}; (X_s)_i \mid (X_t)_{-i}) \cdot \frac{e^{-2h}(1-e^{-s})}{1-e^{-h}+e^{-h}(1-e^{-s})}\,. \tag{8}$$

Finally, by the Markov chain property and then by chain rule,

$$\begin{aligned} I((X_s)_{i+1:d}; (X_s)_i \mid (X_t)_{-i}) &= I((X_s)_{i+1:d}; (X_s)_i, (X_t)_i \mid (X_t)_{-i}) \\ &= I((X_s)_{i+1:d}; (X_s)_i \mid X_t) + I((X_s)_{i+1:d}; (X_t)_i \mid (X_t)_{-i})\,. \end{aligned} \tag{9}$$

Note that the second term on the right-hand side is the same as the left-hand side of Eq. (8). Denoting the factor on the right-hand side of Eq. (8) by C, if we substitute Eq. (9) into Eq. (8) and rearrange, we conclude that

$$I((X_s)_{i+1:d}; (X_s)_i \mid X_t) \leqslant (1/C - 1) \cdot I((X_s)_{i+1:d}; (X_t)_i \mid (X_t)_{-i})\,.$$

Finally, we have

$$1/C - 1 = \frac{(1-e^{-h})(1+e^{-h}(1-e^{-s}))}{e^{-2h}(1-e^{-s})} \leqslant \frac{e^{2h}-1}{1-e^{-s}}\,, \tag{10}$$

as claimed. □

This corollary allows us to establish our key result relating the conditional total correlation $\mathsf{TC}(X_s \mid X_t)$ (in which the terms on the left-hand side of Eq. (5) appear) to the dual total correlation decrement in Lemma 2.4 (in which the terms on the right-hand side of Eq. (5) appear).

Proof of Lemma C.2. By the definition of TC and Corollary C.5,

$$\mathsf{TC}(X_s \mid X_t) = \sum_{i=1}^{d-1} I((X_s)_i; (X_s)_{i+1:d} \mid X_t) \leqslant \frac{e^{2h}-1}{1-e^{-s}} \sum_{i=1}^{d-1} I((X_t)_i; (X_s)_{i+1:d} \mid (X_t)_{-i})\,.$$

Note that

$$I((X_t)_i; (X_s)_{i+1:d} \mid (X_t)_{-i}) \leqslant I((X_t)_i; (X_s)_{-i} \mid (X_t)_{-i})\,,$$

and the sum of these over $i = 1, \ldots, d-1$ is, by Lemma 2.4, at most $\mathsf{DTC}(q_s) - \mathsf{DTC}(q_t)$, as desired. □

Henceforth, for convenience, we denote the factor in Lemma C.2 by

$$c_s(h) \coloneqq \frac{e^{2h}-1}{1-e^{-s}}\,,$$

so that

$$\mathsf{TC}(X_s \mid X_t) \leqslant c_s(h)(\mathsf{DTC}(q_s) - \mathsf{DTC}(q_t))\,.$$

C.2 Score error and approximate reverse kernel

In this section, we show that the approximate score $\widehat{s}$ can be converted into approximate posterior marginals, by giving a "robust" version of Lemma B.1. We then bound the distance between these approximate posterior marginals and the true posterior marginals in terms of the score entropy loss $\varepsilon_{\mathsf{unif}}$.

Given $b \in \Sigma$, define

$$\mathcal{S}_{t,b} \coloneqq \left\{ v \in \mathbb{R}^S_{>0} : v_b = 1 \text{ and } \frac{v_a}{\sum_{a' \in \Sigma} v_{a'}} \geq \beta_t \ \forall\, a \in \Sigma \right\}.$$

As the $S+1$ constraints defining this set are linear, this is a convex set. Furthermore, by the constraint corresponding to $a = b$, the set is compact.

Proposition C.6. *For every $t > 0$, $y \in \mathcal{X}$, and $i \in [d]$, the true score $(s_t(y)[i,a])_{a\in\Sigma}$ is an element of $\mathcal{S}_{t,y_i}$.*

Furthermore, for $0 < s < t$, $h = t - s$, and any $v \in \mathcal{S}_{t,b}$, the distribution $p_h^{\mathsf{gibbs}}[v]$ over Σ defined in Lemma B.1 is a valid probability distribution and satisfies $p_h^{\mathsf{gibbs}}[v](a) \geq \beta_s$ for all $a \in \Sigma$. Likewise, $p_{h,b}^{\mathsf{post}}[v](a) = \left(\sum_{a' \in \Sigma} v_{a'}\right) K_h^{\mathsf{U}}(b \mid a)\, p_h^{\mathsf{gibbs}}[v](a)$ is a valid probability distribution.

We defer the proof to Appendix J.

Let $\mathrm{proj}_{t,b} : \mathbb{R}^S_{>0} \to \mathbb{R}^S_{>0}$ denote the Bregman projection

$$\mathrm{proj}_{t,b}(v) \coloneqq \arg\min_{v' \in \mathcal{S}_{t,b}} \sum_{a \in \Sigma} \psi(v'_a, v_a)\,.$$

This is uniquely defined because $\mathcal{S}_{t,b}$ is convex and compact and ψ is strictly convex in its first argument. By the Pythagorean theorem for Bregman divergences, for any $v^* \in \mathcal{S}_{t,b}$ and any $v \in \mathbb{R}^S_{>0}$,

$$\sum_{a \in \Sigma} \psi(v^*_a, \mathrm{proj}_{t,b}(v)_a) \leqslant \sum_{a \in \Sigma} \psi(v^*_a, v_a)\,. \tag{11}$$

Given score estimates $(\widehat{s}_t)$, define their projections $(\widehat{s}^\sharp_t)$ by

$$\widehat{s}^\sharp_t(y)[i,\cdot] \coloneqq \mathrm{proj}_{t,y_i}(\widehat{s}_t(y)[i,\cdot])\,.$$

For $0 < s < t$ and $h = t - s$, we will consider the approximate reverse kernel

$$\widehat{R}^{\mathsf{unif}}_{t\to s}(y,x) \coloneqq \prod_{i=1}^d \widehat{R}^i_{t\to s}(y, x_i)\,, \qquad \widehat{R}^i_{t\to s}(y,\cdot) \coloneqq p^{\mathsf{post}}_{h,y_i}[\widehat{s}^\sharp_t(y)[i,\cdot]]\,.$$

We will compare this against the true product reverse kernel:

$$\overline{R}^{\mathsf{unif}}_{t\to s}(y,x) \coloneqq \prod_{i=1}^d R^i_{t\to s}(y, x_i)\,, \qquad R^i_{t\to s}(y,\cdot) \coloneqq p^{\mathsf{post}}_{h,y_i}[s_t(y)[i,\cdot]] = \mathbb{P}[(X_s)_i = \cdot \mid X_t = y]\,. \tag{12}$$

Lemma C.7. *For $0 < s < t$ and $h = t - s$,*

$$\mathbb{E}_{Y\sim q_t} \mathsf{KL}(\overline{R}^{\mathsf{unif}}_{t\to s}(Y,\cdot) \parallel \widehat{R}^{\mathsf{unif}}_{t\to s}(Y,\cdot)) \leqslant c_s(h)\varepsilon_{\mathsf{unif}}(t)\,.$$

Proof. Fix any $y \in \mathcal{X}$ and $i \in [d]$ and condition on $(X_t)_{-i} = y_{-i}$. For convenience, denote by p the conditional law of $(X_s)_i$, and also define the corresponding approximation $\widehat{p} = p_h^{\mathsf{gibbs}}[\widehat{s}^\sharp_t(y)[i,\cdot]]$ given by the score estimate. Additionally, given any $b \in \Sigma$, denote by $p_{|b}$ the conditional law of $(X_s)_i$ upon further conditioning on $(X_t)_i = b$. Note that

$$p_{|y_i} = R^i_{t\to s}(y,\cdot)\,,$$

so we would like to bound $\mathsf{KL}(p_{|y_i} \,\|\, \widehat{p}_{|y_i})$ for a suitable approximation $\widehat{p}_{|y_i}$ defined below. Finally, let $r = pK_h^{\mathsf{U}}$ and $\widehat{r} = \widehat{p}K_h^{\mathsf{U}}$, noting that r is the conditional law of $(X_t)_i$ given $(X_t)_{-i} = y_{-i}$, and $\widehat{r}$ is an approximation thereof, so that

$$s_t(y)[i,a] = \frac{r(a)}{r(y_i)} \qquad \text{and} \qquad \widehat{s}_t^{\sharp}(y)[i,a] = \frac{\widehat{r}(a)}{\widehat{r}(y_i)}\,.$$

By chain rule applied to the joint conditional law of $(X_s)_i, (X_t)_i$ and its approximation,

$$\mathsf{KL}(p \,\|\, \widehat{p}) = \mathsf{KL}(r \,\|\, \widehat{r}) + \mathbb{E}_{b\sim r}\mathsf{KL}(p_{|b} \,\|\, \widehat{p}_{|b})\,, \tag{13}$$

where $\widehat{p}_{|b}(\cdot) \coloneqq \widehat{p}(\cdot)K_h^{\mathsf{U}}(b \mid \cdot)/\widehat{r}(b)$ is the posterior law of a sample from $\widehat{p}$ given that its image under the channel K_h^{U} equals b. Writing $v = \widehat{s}_t^{\sharp}(y)[i,\cdot]$, we have $\widehat{r}(a) = \alpha_h\widehat{p}(a) + \beta_h = v_a/\sum_{a'\in\Sigma} v_{a'}$, and substituting this into the definition of $\widehat{p}_{|y_i}$ and using $v_{y_i} = 1$ gives

$$\widehat{p}_{|y_i} = p^{\mathsf{post}}_{h,y_i}[v] = \widehat{R}^i_{t\to s}(y,\cdot)\,.$$

We will lower bound the expectation on the right-hand side of Eq. (13) by the contribution from $b = y_i$ to get

$$\mathsf{KL}(p \,\|\, \widehat{p}) \geq \mathsf{KL}(r \,\|\, \widehat{r}) + r(y_i)\cdot\mathsf{KL}(p_{|y_i} \,\|\, \widehat{p}_{|y_i})\,.$$

By the reverse DPI in Lemma C.4, whose hypothesis holds with $c = 1 - e^{-s}$ by Propositions C.3 and C.6,

$$\mathsf{KL}(r \,\|\, \widehat{r}) \geq \mathsf{KL}(p \,\|\, \widehat{p}) \cdot \frac{e^{-2h}(1-e^{-s})}{1 - e^{-h} + e^{-h}(1-e^{-s})}\,.$$

Recall that this factor is the same one appearing in Eq. (8), which was denoted in the proof of Corollary C.5 by C and which satisfies $1/C - 1 \leqslant c_s(h)$ by Eq. (10). Rearranging, we have

$$\mathsf{KL}(p_{|y_i} \,\|\, \widehat{p}_{|y_i}) \leqslant \frac{c_s(h)}{r(y_i)}\mathsf{KL}(r \,\|\, \widehat{r})\,.$$

We have

$$\begin{aligned}
\sum_a \psi(s_t(y)[i,a], \widehat{s}_t^{\sharp}(y)[i,a]) &= \frac{1}{\widehat{r}(y_i)} - \frac{1}{r(y_i)} + \sum_{a\in\Sigma}\frac{r(a)}{r(y_i)}\log\Big(\frac{r(a)}{\widehat{r}(a)}\cdot\frac{\widehat{r}(y_i)}{r(y_i)}\Big) \\
&= \psi\Big(\frac{1}{r(y_i)}, \frac{1}{\widehat{r}(y_i)}\Big) + \frac{1}{r(y_i)}\mathsf{KL}(r \,\|\, \widehat{r}) \\
&\geq \frac{1}{r(y_i)}\mathsf{KL}(r \,\|\, \widehat{r})\,,
\end{aligned}$$

so

$$\mathsf{KL}(p_{|y_i} \,\|\, \widehat{p}_{|y_i}) \leqslant c_s(h)\sum_a \psi(s_t(y)[i,a], \widehat{s}_t^{\sharp}(y)[i,a]) \leqslant c_s(h)\sum_a \psi(s_t(y)[i,a], \widehat{s}_t(y)[i,a])\,,$$

where in the last step we used the Pythagorean theorem in Eq. (11). Summing over coordinates i, averaging over $y \sim q_t$, and recalling Definition 2.1, we conclude the claimed bound. □

C.3 Sampler and telescoping argument

We consider a sampler of the following form. Fix a grid of times

$$0 < t_0 < t_1 < \cdots < t_M\,.$$

1. Initialize at $\mathrm{Unif}(\mathcal{X})$ at time t_M.
2. For each $j = M-1, \ldots, 0$:
   - Apply $\widehat{R}^{\mathsf{unif}}_{t_{j+1}\to t_j}$ to get the next iterate.
3. Given the iterate Y at time t_0, output a draw from $(K_{t_0}^{\mathsf{U}})^{\otimes d}(\cdot \mid Y)$.

The following lemma provides a straightforward chain rule calculation that utilizes the main estimates from the previous subsections (Lemma C.2 and Lemma C.7).

Lemma C.8. *Let $\widehat{q}$ denote the output law of the above sampler. Then*

$$\mathsf{KL}(q \parallel \widehat{q}) \leqslant \mathsf{KL}(q_{t_M} \parallel \mathrm{Unif}(\mathcal{X})) + H(q_{t_0}) - H(q) + \sum_{j=0}^{M-1} c_{t_j}(t_{j+1} - t_j) \cdot \left[\mathsf{DTC}(q_{t_j}) - \mathsf{DTC}(q_{t_{j+1}}) + \varepsilon_{\mathsf{unif}}(t_{j+1})\right].$$

Proof. Let P be the joint law of $(X_{t_M}, X_{t_{M-1}}, \dots, X_{t_0}, X_0)$ under the forward process, written in reverse order. Let $\widehat{P}$ be the comparison path law that initializes its first coordinate from $\mathrm{Unif}(\mathcal{X})$, uses $\widehat{R}^{\mathsf{unif}}_{t_{j+1}\to t_j}$ from t_{j+1} to t_j, and uses $(K^{\mathsf{U}}_{t_0})^{\otimes d}(\cdot \mid X_{t_0})$ for the last transition. For $0 \leq j < M$, define

$$\begin{aligned} E_j^{\mathrm{disc}} &:= \mathbb{E}_{X_{t_{j+1}}} \mathsf{KL}(\mathrm{law}(X_{t_j} \mid X_{t_{j+1}}) \parallel \overline{R}^{\mathsf{unif}}_{t_{j+1}\to t_j}(X_{t_{j+1}}, \cdot)), \\ E_j^{\mathrm{sc}} &:= \mathbb{E}_{X_{t_{j+1}}} \mathsf{KL}(\overline{R}^{\mathsf{unif}}_{t_{j+1}\to t_j}(X_{t_{j+1}}, \cdot) \parallel \widehat{R}^{\mathsf{unif}}_{t_{j+1}\to t_j}(X_{t_{j+1}}, \cdot)), \\ E_0^{\mathrm{end}} &:= \mathbb{E}_{X_{t_0}} \mathsf{KL}(\mathrm{law}(X_0 \mid X_{t_0}) \parallel (K^{\mathsf{U}}_{t_0})^{\otimes d}(\cdot \mid X_{t_0})). \end{aligned}$$

Because the marginals of the product reverse kernel $\overline{R}^{\mathsf{unif}}_{t_{j+1}\to t_j}$ are by definition the marginals of $\mathrm{law}(X_{t_j} \mid X_{t_{j+1}} = y)$,

$$\mathbb{E}_{X_{t_{j+1}}} \mathsf{KL}(\mathrm{law}(X_{t_j} \mid X_{t_{j+1}}) \parallel \widehat{R}^{\mathsf{unif}}_{t_{j+1}\to t_j}(X_{t_{j+1}}, \cdot)) = E_j^{\mathrm{disc}} + E_j^{\mathrm{sc}}.$$

The X_0 marginal of P is q, while the last-coordinate marginal of $\widehat{P}$ is $\widehat{q}$. Data processing and chain rule for KL therefore give

$$\begin{aligned} \mathsf{KL}(q \parallel \widehat{q}) &\leq \mathsf{KL}(P \parallel \widehat{P}) \\ &= \mathsf{KL}(q_{t_M} \parallel \mathrm{Unif}(\mathcal{X})) + \sum_{j=0}^{M-1} (E_j^{\mathrm{disc}} + E_j^{\mathrm{sc}}) + E_0^{\mathrm{end}}. \end{aligned}$$

By Lemma C.9 below, the last term is equal to $H(q_{t_0}) - H(q)$. Furthermore, by definition $E_j^{\mathrm{disc}} = \mathsf{TC}(X_{t_j} \mid X_{t_{j+1}})$, so by Lemma C.2,

$$E_j^{\mathrm{disc}} \leqslant c_{t_j}(t_{j+1} - t_j)\big(\mathsf{DTC}(q_{t_j}) - \mathsf{DTC}(q_{t_{j+1}})\big).$$

Finally, by Lemma C.7,

$$E_j^{\mathrm{sc}} \leqslant c_{t_j}(t_{j+1} - t_j)\varepsilon_{\mathsf{unif}}(t_{j+1}).$$

Substituting all of these bounds proves the claim. □

The lemma below was used above to control the error incurred by the rounding step at the end of the sampler.

Lemma C.9. *For any $t > 0$, we have*

$$\mathbb{E}_{X_t \sim q_t} \mathsf{KL}(\mathrm{law}(X_0 \mid X_t) \parallel (K^{\mathsf{U}}_t)^{\otimes d}(\cdot \mid X_t)) = H(q_t) - H(q).$$

Proof. The joint law of (X_0, X_t) is $q(x)(K^{\mathsf{U}}_t)^{\otimes d}(y \mid x)$, while the comparison joint law is $q_t(y)(K^{\mathsf{U}}_t)^{\otimes d}(x \mid y)$. The kernel is symmetric, so their log-likelihood ratio is $\log q(x) - \log q_t(y)$. Averaging gives $-H(q) + H(q_t)$. □

We next control the two endpoint terms in Lemma C.8 by taking t_0 sufficiently small and t_M sufficiently large.

Lemma C.10. *For any $T > 0$, $\mathsf{KL}(q_T \parallel \mathrm{Unif}(\mathcal{X})) \leqslant e^{-T} d \log S$.*

We defer the proof to Appendix J.

Lemma C.11. *For $0 < t \leqslant 1$, $H(q_t) - H(q) \leqslant dt \log(eS/t)$.*

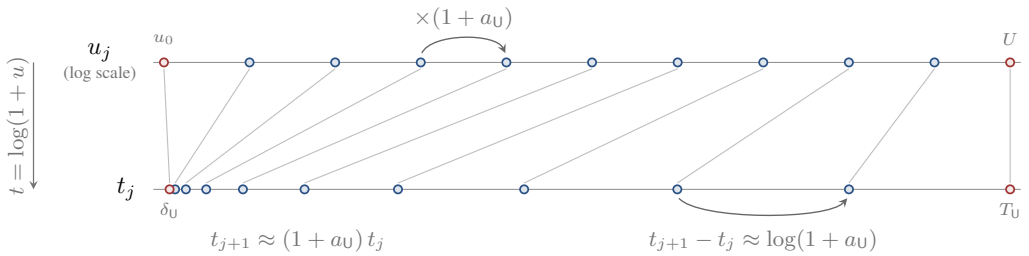


Figure 5: Step schedule for uniform diffusion sampler

We defer the proof to Appendix J.

We are now ready to prove our main result, a DTC-adaptive query complexity for uniform diffusion sampling:

Proof of Theorem C.1. It remains to set the step sizes so that the sum in Lemma C.8 telescopes (see Figure 5). Define

$$L_{\mathsf{U}} = \log(16edS/\varepsilon), \qquad \delta_{\mathsf{U}} = \frac{\varepsilon}{16dL_{\mathsf{U}}}, \qquad T_{\mathsf{U}} = \log(4d\log(S)/\varepsilon)\,,$$

and

$$a_{\mathsf{U}} = \frac{\varepsilon}{12\max\{\overline{\mathsf{DTC}},\varepsilon\}}, \qquad u_0 = e^{\delta_{\mathsf{U}}}-1, \qquad U = e^{T_{\mathsf{U}}}-1\,,$$

noting that $a_{\mathsf{U}} \leqslant 1/12$. Recursively set

$$u_{j+1} = \min((1+a_{\mathsf{U}})u_j, U)\,,$$

and let N_{U} be the first index for which $u_j = U$. Set $t_j = \log(1+u_j)$, so that $\delta_{\mathsf{U}} = t_0 < \cdots < t_{N_{\mathsf{U}}} = T_{\mathsf{U}}$. Because $u_j/(1+u_j) = 1-e^{-t_j}$, we have

$$e^{t_{j+1}-t_j} = \frac{1+u_{j+1}}{1+u_j} \leqslant \frac{1+(1+a_{\mathsf{U}})u_j}{1+u_j} = 1+a_{\mathsf{U}}(1-e^{-t_j})\,,$$

and hence, writing $x_j \coloneqq a_{\mathsf{U}}(1-e^{-t_j}) \leqslant 1$,

$$c_{t_j}(t_{j+1}-t_j) = \frac{e^{2(t_{j+1}-t_j)}-1}{1-e^{-t_j}} \leqslant \frac{(1+x_j)^2-1}{1-e^{-t_j}} = \frac{(2+x_j)x_j}{1-e^{-t_j}} \leqslant 3a_{\mathsf{U}}\,.$$

Therefore,

$$\sum_{j=0}^{N_{\mathsf{U}}-1} c_{t_j}(t_{j+1}-t_j)\big(\mathsf{DTC}(q_{t_j}) - \mathsf{DTC}(q_{t_{j+1}}) + \varepsilon_{\mathsf{unif}}(t_{j+1})\big) \leq 3a_{\mathsf{U}}\mathsf{DTC}(q) + 3a_{\mathsf{U}}N_{\mathsf{U}}\varepsilon_{\mathsf{unif}}$$
$$\leqslant \varepsilon/4 + 3a_{\mathsf{U}}N_{\mathsf{U}}\varepsilon_{\mathsf{unif}}\,,$$

since $\mathsf{DTC}(q_t)$ is nonincreasing in t by Lemma 2.4, so the increments are nonnegative and telescope.

By Lemma C.10 and the choice of T_{U}, $\mathsf{KL}(q_{T_{\mathsf{U}}} \,\|\, \mathrm{Unif}(\mathcal{X})) \leq \varepsilon/4$. By Lemma C.11 and the choice of δ_{U}, $H(q_{\delta_{\mathsf{U}}}) - H(q) \leqslant d\delta_{\mathsf{U}}\log(eS/\delta_{\mathsf{U}}) = \frac{\varepsilon}{16}\big(1+\frac{\log L_{\mathsf{U}}}{L_{\mathsf{U}}}\big) \leqslant \varepsilon/8$.

Finally, u_j increases by a factor $1+a_{\mathsf{U}}$ until the last step, so

$$N_{\mathsf{U}} \leq 1 + \frac{\log(U/u_0)}{\log(1+a_{\mathsf{U}})}\,.$$

Here $\log(U/u_0) \lesssim L_{\mathsf{U}}$ and $1/\log(1+a_{\mathsf{U}}) \leq 2/a_{\mathsf{U}} \lesssim 1+\overline{\mathsf{DTC}}/\varepsilon$. Each application of $\widehat{R}^{\mathsf{unif}}_{t_{j+1}\to t_j}$ requires only the score matrix $\widehat{s}_{t_{j+1}}(y) \in \mathbb{R}^{[d]\times\Sigma}$ at the current iterate y, i.e., one score oracle query, so the sampler makes N_{U} queries in total, which proves the query bound.

Substituting all of the above bounds into Lemma C.8 gives the claimed KL bound. □

D GAUSSIAN DIFFUSION CAN SCALE WITH DUAL TOTAL CORRELATION

We now prove the Gaussian diffusion analogue of Theorem C.1. Recall from Section 2.2 that we consider the pushforward q of a discrete distribution q_{pre} on $\mathcal{X} = \Sigma^d$ to $(\mathbb{R}^S)^d$ via the one-hot encoding. When we refer to the dual total correlation of q, we treat each block of $S = |\Sigma|$ entries as a single coordinate in the definition of $\mathsf{DTC}(q)$, so that $\mathsf{DTC}(q) = \mathsf{DTC}(q_{\mathsf{pre}})$.

Given $X_0 \sim q$, let (X_t) be the Ornstein–Uhlenbeck process of Section 2.2, with $q_t := \mathrm{law}(X_t)$ and score $s_t = \nabla \log q_t$. The error of estimated scores $(\widehat{s}_t)$ is quantified by the L^2 score error $\varepsilon_{\mathsf{gauss}}(t)$ of Definition 2.2.

Theorem D.1. *Let q be any one-hot encoding of a distribution q_{pre} on $\mathcal{X}$, and suppose that score estimates $(\widehat{s}_t)$ satisfying $\varepsilon_{\mathsf{gauss}}(t) \leqslant \varepsilon_{\mathsf{gauss}}$ for all t, along with a number $\overline{\mathsf{DTC}} \geq \mathsf{DTC}(q)$, are available. For every $0 < \varepsilon < 1$, there is a sampler whose output law $\widehat{q}$ on $[S]^d$ satisfies*

$$\mathsf{KL}(q_{\mathsf{pre}} \,\|\, \widehat{q}) \leqslant \varepsilon + O(\varepsilon^2_{\mathsf{gauss}} \cdot \log(dS/\varepsilon)) \,,$$

and which uses

$$O\left(\left(1 + \frac{\overline{\mathsf{DTC}}}{\varepsilon}\right) \log \frac{edS}{\varepsilon}\right)$$

score oracle queries.

The general outline of the argument remains the same as the one for uniform diffusion in Section C. An important difference is that there is no analogue of Proposition C.3, so the reverse DPI used to relate conditional TC to the decrease in DTC must be proven using a different route (see Section D.2).

Change of variable. Throughout this section, in place of the OU parametrization it will be more convenient to work with the additive noise parametrization. As such, we consider the change of variable

$$u = e^{2t} - 1 \,, \qquad t(u) = \tfrac{1}{2}\log(1+u) \,, \qquad Z_u := e^{t(u)} X_{t(u)} \,,$$

so that

$$Z_u = X_0 + B_u \tag{14}$$

for a standard Brownian motion $(B_u)_{u \geq 0}$ in $\mathbb{R}^{Sd}$ independent of X_0.

In lieu of q_t and s_t, we use π_u and $g_u = \nabla \log \pi_u$ to denote the density of Z_u and its score, noting that

$$g_u(z) := \frac{1}{\sqrt{1+u}} s_{t(u)}\left(\frac{z}{\sqrt{1+u}}\right).$$

Similarly, we will consider the score estimate

$$\widehat{g}_u(z) := \frac{1}{\sqrt{1+u}} \widehat{s}_{t(u)}\left(\frac{z}{\sqrt{1+u}}\right),$$

which has score error

$$\begin{aligned}\mathbb{E}_{Z_u \sim \pi_u} \|\widehat{g}_u(Z_u) - g_u(Z_u)\|_2^2 &= \frac{1}{1+u}\, \mathbb{E}_{X_{t(u)} \sim q_{t(u)}} \|\widehat{s}_{t(u)}(X_{t(u)}) - s_{t(u)}(X_{t(u)})\|_2^2 \\ &= \frac{1+u}{u^2}\, \varepsilon_{\mathsf{gauss}}(t(u))^2 \,,\end{aligned} \tag{15}$$

where in the last step we re-expressed the $\sigma^2_{t(u)}$ prefactor in the definition of $\varepsilon_{\mathsf{gauss}}$ via $\sigma^2_{t(u)} = u/(1+u)$.

D.1 CONDITIONAL TC VERSUS DTC DECREMENT

The main result of this subsection is the following analogue of Lemma C.2: an upper bound on the conditional TC, $\mathsf{TC}(Z_u \mid Z_v)$, in terms of the amount by which the dual total correlation decreases when going from noise level u to higher noise level v. Here and below, for laws with densities DTC is defined by the same formula with differential entropies, and Lemma 2.4 applies verbatim (also when A is discrete and B continuous).

Lemma D.2. *For* $0 < u < v$,

$$\mathsf{TC}(Z_u \mid Z_v) \leqslant \big(\frac{v}{u}e^{3/u-3/v} - 1\big)\big(\mathsf{DTC}(\pi_u) - \mathsf{DTC}(\pi_v)\big)\,.$$

As in the uniform diffusion case, the key ingredient in the proof is a reverse DPI for a single token. In the uniform diffusion case, however, the proof of the reverse DPI crucially relied on a lower bound on the mass that the relevant conditional distributions placed on each possible token. In the Gaussian setting we cannot rely on such a property.

For a distribution $\alpha = (\alpha_1, \ldots, \alpha_S)$ over Σ and $u > 0$, if we regard it as a distribution over standard basis vectors in $\mathbb{R}^S$ and apply Gaussian noise at level u, the resulting distribution is a mixture of Gaussians with density

$$f_{\alpha,u}(z) = \sum_{a\in\Sigma} \alpha_a \phi_u(z - e_a)\,,$$

where ϕ_u denotes the density of $N(0, u \cdot \mathrm{Id}_S)$.

The main estimate we show is the following reverse DPI for such mixture distributions.

Lemma D.3 (Reverse DPI for Gaussian diffusion)**.** *For every* $\alpha, \beta \in \Delta(\Sigma)$ *and* $0 < u < v$,

$$\mathsf{KL}(f_{\alpha,v} \parallel f_{\beta,v}) \geq \mathsf{KL}(f_{\alpha,u} \parallel f_{\beta,u}) \cdot \Big(\frac{u}{v}e^{3/v-3/u}\Big)\,.$$

Lemma D.3 is a consequence of standard identities for the additive Gaussian channel, and we defer its proof to Section D.2 and first derive Lemma D.2 from it.

In exact analogy with Corollary C.5 in the uniform diffusion case, we use this bound to compare two quantities: (1) the mutual information between a block i of Z_u and the later blocks of Z_u, conditioned on the sequence Z_v at a higher noise level v, and (2) the same quantity except $Z_u^{(i)}$ is replaced with $Z_v^{(i)}$ and only the remaining blocks of Z_v are conditioned upon (recall Figure 4).

Corollary D.4. *For* $0 < u < v$ *and* $i \in [d-1]$,

$$I(Z_u^{(i+1:d)}; Z_u^{(i)} \mid Z_v) \leqslant \big(\frac{v}{u}e^{3/u-3/v} - 1\big) \cdot I(Z_u^{(i+1:d)}; Z_v^{(i)} \mid Z_v^{(-i)})\,.$$

Proof. Conditionally on $(Z_u^{(i+1:d)}, Z_v^{(-i)})$, the block $Z_u^{(i)}$ has law $f_{\alpha,u}$ and $Z_v^{(i)}$ has law $f_{\alpha,v}$ with $\alpha = \mathrm{law}(A_i \mid Z_u^{(i+1:d)}, Z_v^{(-i)})$, because the noise in block i is independent of the conditioning; likewise with $\beta = \mathrm{law}(A_i \mid Z_v^{(-i)})$ when conditioning on $Z_v^{(-i)}$ alone. Lemma D.3 therefore applies to these conditional laws, and the proof follows verbatim from the argument for Corollary C.5 in the uniform diffusion case, with the parameter C defined therein taken to be $C = \frac{u}{v}e^{3/v-3/u}$ in light of Lemma D.3. □

Proof of Lemma D.2. The proof also follows verbatim from the argument for Lemma C.2, but where $\frac{e^{2h}-1}{1-e^{-s}}$ therein is replaced with $\frac{v}{u}e^{3/u-3/v} - 1$. □

Henceforth, for convenience denote the factor in Lemma D.2 by

$$c_u(v) \coloneqq \frac{v}{u}e^{3/u-3/v} - 1\,,$$

so that

$$\mathsf{TC}(Z_u \mid Z_v) \leqslant c_u(v)(\mathsf{DTC}(\pi_u) - \mathsf{DTC}(\pi_v))\,.$$

D.2 Reverse DPI for Gaussian diffusion

We will use the following standard identities for the additive Gaussian channel, see, e.g., (Wibisono, 2025, Lemmas 1 and 2):

Lemma D.5. *Let $(\mu_u)_{u\geq 0}$ and $(\nu_u)_{u\geq 0}$ be measures evolving according to the standard Gaussian channel, that is, which satisfy $\partial_u p_u = \frac{1}{2}\Delta p_u$ for $p = \mu, \nu$. Denote the log-density ratio between them by $\ell_u \coloneqq \log(\mu_u/\nu_u)$, and denote the* relative Fisher information *between them by*

$$\mathsf{FI}(\mu_u \parallel \nu_u) \coloneqq \mathbb{E}_{\mu_u}\|\nabla \ell_u\|_2^2 \,.$$

Then

- $\partial_u \mathsf{KL}(\mu_u \parallel \nu_u) = -\frac{1}{2}\mathsf{FI}(\mu_u \parallel \nu_u)$
- $\partial_u \mathsf{FI}(\mu_u \parallel \nu_u) = -\mathbb{E}_{\mu_u}\|\nabla^2 \ell_u\|_F^2 + 2\mathbb{E}_{\mu_u}(\nabla \ell_u)^\intercal(\nabla^2 \log \nu_u)(\nabla \ell_u)$

Specializing μ_u and ν_u above to $f_{\alpha,u}$ and $f_{\beta,u}$ respectively, we obtain the following:

Lemma D.6. *Let $\alpha, \beta \in \Delta(\Sigma)$ and $u > 0$. Given $\lambda \in \Delta(\Sigma)$, define*

$$\mathsf{m}_{\lambda,u}(z) \coloneqq \left(\frac{\lambda_a e^{z_a/u}}{\sum_b \lambda_b e^{z_b/u}}\right)_{a\in\Sigma} \qquad \textit{and} \qquad \mathsf{C}_{\lambda,u}(z) \coloneqq \operatorname{diag}(\mathsf{m}_{\lambda,u}(z)) - \mathsf{m}_{\lambda,u}(z)^{\otimes 2} \,.$$

Then

- $\partial_u \mathsf{KL}(f_{\alpha,u} \parallel f_{\beta,u}) = -\frac{1}{2}\mathsf{FI}(f_{\alpha,u} \parallel f_{\beta,u})$
- $\partial_u \mathsf{FI}(f_{\alpha,u} \parallel f_{\beta,u}) \geq -\left(\frac{2}{u} + \frac{3}{u^2}\right)\mathsf{FI}(f_{\alpha,u} \parallel f_{\beta,u})$

We defer the proof to Appendix J.

We are now ready to complete the proof of Lemma D.3.

Proof of Lemma D.3. For convenience, define $K(u) \coloneqq \mathsf{KL}(f_{\alpha,u} \parallel f_{\beta,u})$. By Lemma D.6, we have

$$-K'(u) = \frac{1}{2}\mathsf{FI}(f_{\alpha,u} \parallel f_{\beta,u}) \geq 0 \quad \text{and} \quad K''(u) = -\frac{1}{2}\partial_u \mathsf{FI}(f_{\alpha,u} \parallel f_{\beta,u}) \leqslant -\left(\frac{2}{u} + \frac{3}{u^2}\right)K'(u) \,.$$

Let $\xi = v/u > 1$. By Grönwall's inequality applied to $-K'$, for all $s \geq u$ we have

$$-K'(\xi s) \geq -K'(s)\exp\left(-\int_s^{\xi s}\left(\frac{2}{r} + \frac{3}{r^2}\right)\mathrm{d}r\right) = -K'(s)\frac{1}{\xi^2}e^{3/(\xi s) - 3/s} \geq -K'(s)\frac{1}{\xi^2}e^{3/v - 3/u} \,.$$

Note that $K(\infty) = 0$, so

$$K(u) = -\int_u^\infty K'(s)\,\mathrm{d}s \leqslant -e^{3/u - 3/v}\int_u^\infty \xi^2 K'(\xi s)\,\mathrm{d}s = \frac{v}{u}e^{3/u - 3/v}K(v) \,,$$

as claimed. □

D.3 Score error and approximate reverse kernel

Note that from the true score we can read off information about the posterior per-token marginals. Indeed, by Tweedie's formula (Lemma B.3),

$$m_v(z) \coloneqq \mathbb{E}[X_0 \mid Z_v = z] = z + v g_v(z) \,,$$

and for every block $i \in [d]$,

$$m_v^{(i)}(z) = \mathbb{P}[A_i = \cdot \mid Z_v = z] \,. \tag{16}$$

Conditioned on $(X_0, Z_v = z)$, the bridge law of the forward process in Eq. (14) is

$$Z_u \sim N\left(\frac{u}{v}z + \left(1 - \frac{u}{v}\right)X_0,\ u\left(1 - \frac{u}{v}\right)\mathrm{Id}_{Sd}\right) \,.$$

Therefore, the reverse kernel's marginals are given by a mixture of Gaussians:

$$\begin{aligned} R_{v\to u}^{(i)}(y, \cdot) &\coloneqq \operatorname{law}\left(Z_u^{(i)} \mid Z_v = y\right) \\ &= \sum_{a\in\Sigma} m_{v,a}^{(i)}(y) \cdot N\left(\frac{u}{v}y^{(i)} + \left(1 - \frac{u}{v}\right)e_a,\ u\left(1 - \frac{u}{v}\right)\mathrm{Id}_S\right) \,. \end{aligned}$$

To sample from $R^{(i)}_{v\to u}(y,\cdot)$, one draws $\xi_i = e_a$ with probability $m^{(i)}_{v,a}(y)$ and outputs $\frac{u}{v}y^{(i)} + (1 - \frac{u}{v})\xi_i + \sqrt{u(1-\frac{u}{v})}\,G_i$, where $G_i \sim N(0, \mathrm{Id}_S)$ is drawn independently across blocks. This naturally suggests the following product kernel in analogy with Eq. (12):

$$\overline{R}^{\mathsf{G}}_{v\to u}(y,\cdot) \coloneqq \prod_{i=1}^{d} R^{(i)}_{v\to u}(y,\cdot)\,.$$

In this subsection, we show that the approximate score $\widehat{g}$ can likewise give rise to an approximation of this product kernel. We then bound the distance between the approximation and the true product kernel in terms of the score error $\varepsilon_{\mathsf{gauss}}$.

In lieu of the Bregman projection from Section C.2, here we simply use a Euclidean projection $\Pi_{\Delta(\Sigma)^d}$: given the rescaled estimate $\widehat{g}_v$, define

$$\widehat{m}_v(z) \coloneqq \Pi_{\Delta(\Sigma)^d}\big(z + v\widehat{g}_v(z)\big)\,. \tag{17}$$

Since $m_v(z) \in \Delta(\Sigma)^d$ by Eq. (16) and Euclidean projection onto the closed convex set $\Delta(\Sigma)^d$ is nonexpansive, we have

$$\|\widehat{m}_v(z) - m_v(z)\|_2 \le v\|\widehat{g}_v(z) - g_v(z)\|_2\,, \tag{18}$$

in analogy with the Pythagorean theorem in Eq. (11). We can then construct the approximate reverse kernel

$$\widehat{R}^{\mathsf{G}}_{v\to u}(y,\cdot) \coloneqq \prod_{i=1}^{d} \widehat{R}^{(i)}_{v\to u}(y,\cdot)\,,$$

by replacing m_v above with $\widehat{m}_v$ to obtain

$$\widehat{R}^{(i)}_{v\to u}(y,\cdot) \coloneqq \sum_{a\in\Sigma} \widehat{m}^{(i)}_{v,a}(y)\cdot N\Big(\frac{u}{v}y^{(i)} + \Big(1-\frac{u}{v}\Big)e_a,\; u\Big(1-\frac{u}{v}\Big)\mathrm{Id}_S\Big)\,.$$

We now proceed to bound the distance between the product kernel and its approximation. To do so, we would like to quantify the distance between two mixtures of Gaussians at a given noise level in terms of the squared Euclidean distance between their mixing coefficients:

Lemma D.7. *For $\alpha, \beta \in \Delta(\Sigma)$ and $\rho > 0$,*

$$\mathsf{KL}(f_{\alpha,\rho} \parallel f_{\beta,\rho}) \le \big(e^{2/\rho} - 1\big)\|\alpha - \beta\|_2^2\,.$$

We defer the proof to Appendix J.

From this, we can readily control the distance between the product kernel and its approximation in terms of the score estimation error:

Lemma D.8. *For $0 < u < v$, let*

$$w_u(v) \coloneqq (1+v)\left[\exp\left(\frac{2(v-u)}{uv}\right) - 1\right]\,.$$

Then

$$\mathbb{E}_{y\sim\pi_v}\mathsf{KL}(\overline{R}^{\mathsf{G}}_{v\to u}(y,\cdot) \parallel \widehat{R}^{\mathsf{G}}_{v\to u}(y,\cdot)) \leqslant w_u(v)\,\varepsilon_{\mathsf{gauss}}(t(v))^2\,.$$

Moreover, if $c_u(v) \leqslant 1$, then

$$w_u(v) \le 8\log(1 + c_u(v))\,. \tag{19}$$

Proof. As both kernels are products over blocks, it suffices to bound the KL for each block. Over a fixed block $i \in [d]$, the kernels are given by mixtures of Gaussians. After translating both by $(u/v)y^{(i)}$ and scaling by $(1-u/v)^{-1}$, these mixtures become the mixtures $f_{m^{(i)}_v(y),\rho}$ and $f_{\widehat{m}^{(i)}_v(y),\rho}$ from Lemma D.7, with $\rho = uv/(v-u)$. By Lemma D.7 and Eq. (18),

$$\sum_{i=1}^{d}\mathsf{KL}(R^{(i)}_{v\to u}(y,\cdot) \parallel \widehat{R}^{(i)}_{v\to u}(y,\cdot)) \le \big(e^{2(v-u)/(uv)} - 1\big)v^2\|\widehat{g}_v(y) - g_v(y)\|_2^2\,.$$

Averaging over $y \sim \pi_v$ and applying Eq. (15) gives the first part of the claim.

For Eq. (19), write $\theta \coloneqq (v-u)/(uv)$. The hypothesis implies $v/u \le 2$ and $3\theta \le \log 2$, hence $e^{2\theta} - 1 \le 4\theta$ and also $(v-u)/u \le 2\log(v/u)$. It follows that

$$w_u(v) \le 4\theta + \frac{4(v-u)}{u} \le 4\left(\frac{1}{u} - \frac{1}{v}\right) + 8\log\frac{v}{u} \le 8\log(1 + c_u(v))\,. \qquad \square$$

D.4 Sampler and telescoping argument

We consider a sampler of the following form. Fix a grid of noise levels

$$0 < u_0 < u_1 < \cdots < u_M .$$

1. Initialize at $N(0, u_M \mathrm{Id}_{Sd})$ at noise level u_M.
2. For each $j = M-1, \ldots, 0$:
   - Apply $\widehat{R}^{\mathsf{G}}_{u_{j+1} \to u_j}$ to get the next iterate.
3. Given the iterate Z at the smallest noise level u_0, round it to a string in $\mathcal{X}$.

The rounding in Step 3 proceeds as follows. In the uniform diffusion case, recall that we simply applied the forward kernel coordinatewise. In the Gaussian diffusion case, we will instead round the final iterate to a string in $\mathcal{X}$ by drawing from the product of block-wise posterior marginals (under uniform prior) at $u = u_0$:

$$Q_u(a_1, \ldots, a_d \mid z) \coloneqq \prod_{i=1}^{d} Q_u^{(i)}(a_i \mid z^{(i)}), \qquad Q_u^{(i)}(a \mid z^{(i)}) \coloneqq \frac{\exp(z_a^{(i)}/u)}{\sum_{b \in \Sigma} \exp(z_b^{(i)}/u)} .$$

We have the following analogue of the chain rule calculation from Lemma C.8 which combines the main estimates from the previous subsections (Lemma D.2 and Lemma D.8).

Lemma D.9. *Let $\widehat{q}$ denote the output law of the above sampler. Then*

$$\begin{aligned} \mathsf{KL}(q_{\mathsf{pre}} \parallel \widehat{q}) \leqslant \mathsf{KL}(\pi_{u_M} \parallel N(0, u_M \mathrm{Id}_{Sd})) + \mathbb{E}_{Z_{u_0} \sim \pi_{u_0}} \mathsf{KL}(\mathrm{law}(A \mid Z_{u_0}) \parallel Q_{u_0}(\cdot \mid Z_{u_0})) \\ + \sum_{j=0}^{M-1} \Big[c_{u_j}(u_{j+1}) \big(\mathsf{DTC}(\pi_{u_j}) - \mathsf{DTC}(\pi_{u_{j+1}}) \big) + w_{u_j}(u_{j+1})\, \varepsilon_{\mathsf{gauss}}(t(u_{j+1}))^2 \Big] . \end{aligned}$$

Proof. The proof follows verbatim from the argument for Lemma C.8, with the reversed forward path $(X_{t_M}, \ldots, X_{t_0}, X_0)$ replaced by $(Z_{u_M}, \ldots, Z_{u_0}, A)$, the initialization $\mathrm{Unif}(\mathcal{X})$ by $N(0, u_M \mathrm{Id}_{Sd})$, the kernel $(K^{\mathsf{U}}_{t_0})^{\otimes d}$ by Q_{u_0}, and Lemmas C.2 and C.7 by Lemmas D.2 and D.8. The error from the rounding step $\mathbb{E}_{Z_{u_0} \sim \pi_{u_0}} \mathsf{KL}(\mathrm{law}(A \mid Z_{u_0}) \parallel Q_{u_0}(\cdot \mid Z_{u_0}))$ will be analyzed in a subsequent lemma. □

We next control the initialization and rounding errors in Lemma D.9 by taking u_M sufficiently large and u_0 sufficiently small, respectively; the following two results play the roles of Lemmas C.10 and C.11.

Lemma D.10. *For every $U > 0$,*

$$\mathsf{KL}(\pi_U \parallel N(0, U \mathrm{Id}_{Sd})) \leq \frac{d}{2U} .$$

Proof. By convexity of KL in its first argument, it suffices to compare $N(x, U\mathrm{Id}_{Sd})$ with $N(0, U\mathrm{Id}_{Sd})$ for a fixed x whose blocks are standard basis vectors. As the two covariances agree, the KL is simply $\|x\|_2^2/(2U) = d/(2U)$. □

Lemma D.11. *For $u > 0$, let*

$$p^*(u) \coloneqq \frac{S-1}{2} \mathrm{erfc}\Big(\frac{1}{2\sqrt{u}} \Big) .$$

If $p^*(u) \leq 1/2$, then

$$\mathbb{E}_{Z_u \sim \pi_u} \mathsf{KL}(\mathrm{law}(A \mid Z_u) \parallel Q_u(\cdot \mid Z_u)) \leq d\big(h_2(p^*(u)) + p^*(u) \log(S-1) \big) .$$

Proof. The left-hand side is at most $\mathbb{E}[-\log Q_u(A \mid Z_u)]$, by nonnegativity of entropy. As Q_u is a product kernel, this decomposes into a sum $\sum_{i=1}^{d} \mathbb{E}[-\log Q_u^{(i)}(A_i \mid Z_u^{(i)})]$.

Fix any block $i \in [d]$ and $a \in \Sigma$, and consider the contribution of this block conditioned on $A_i = a$, namely

$$\mathbb{E}[-\log Q_u^{(i)}(A_i \mid Z_u^{(i)}) \mid A_i = a]\,. \tag{20}$$

This has the following interpretation. Nature samples A' uniformly at random from Σ and $\gamma \sim N(0, u\mathrm{Id}_S)$, and one observes the vector $z = e_{A'} + \gamma$. Then $Q_u^{(i)}(a \mid z)$ is the observer's posterior probability that $A' = a$. Conditional on $A' = a$, we have $z = e_a + \gamma$, which has the same distribution as $Z_u^{(i)}$ conditional on $A_i = a$, so in Eq. (20) the expectation is over the randomness of γ. As a result, Eq. (20) is independent of a. So we may freely replace the expectation over the marginal distribution of A_i under $A \sim q_{\mathsf{pre}}$ in $\mathbb{E}[-\log Q_u^{(i)}(A_i \mid Z_u^{(i)})]$ with an expectation over any distribution, in particular, over $A_i \sim \mathrm{Unif}(\Sigma)$. Thus, $\mathbb{E}[-\log Q_u^{(i)}(A_i \mid Z_u^{(i)})]$ is nothing more than the conditional entropy of a uniformly random element A' of Σ conditioned on observing $e_{A'} + \gamma$. Consider any decoder $F : \mathbb{R}^S \to \Sigma$ that tries to predict the former given the latter, and let $p_{\mathrm{err}} \coloneqq \mathbb{P}[F(e_{A'} + \gamma) \neq A']$ denote the decoding error.

By Fano's inequality, the conditional entropy is at most

$$h_2(p_{\mathrm{err}}) + p_{\mathrm{err}} \log(S-1)\,,$$

where h_2 denotes binary entropy. Take the decoder F to simply choose the largest coordinate of $e_{A'} + \gamma$. For any realization $A' = a$, the decoding error is the probability that $\gamma_b - \gamma_a > 1$ for some $b \neq a$, which by a union bound is at most $\frac{S-1}{2}\mathrm{erfc}\big(1/(2\sqrt{u})\big)$. Since $p_{\mathrm{err}} \leqslant p^*(u) \leqslant 1/2$ and $p \mapsto h_2(p) + p\log(S-1)$ is increasing on $[0, 1/2]$, the claimed bound then follows by summing over $i \in [d]$. □

We are now ready to prove our main result, a DTC-adaptive query complexity for Gaussian diffusion:

Proof of Theorem D.1. It remains to set the noise levels so that the sum in Lemma D.9 telescopes (see Figure 6). Define

$$L_{\mathsf{G}} = \log(64edS/\varepsilon), \qquad u_0 = \frac{1}{8L_{\mathsf{G}}}, \qquad U = \max\left\{1, \frac{4d}{\varepsilon}\right\},$$

and

$$a_{\mathsf{G}} = \frac{\varepsilon}{4\max\{\overline{\mathsf{DTC}}, \varepsilon\}}, \qquad F_{\mathsf{G}}(u) = \log u - \frac{3}{u}\,,$$

noting that $a_{\mathsf{G}} \leqslant 1/4$. Recursively set u_{j+1} by

$$F_{\mathsf{G}}(u_{j+1}) - F_{\mathsf{G}}(u_j) = \min\left\{\log(1 + a_{\mathsf{G}}),\, F_{\mathsf{G}}(U) - F_{\mathsf{G}}(u_j)\right\}\,,$$

and let N_{G} be the first index for which $u_j = U$; this is well defined because F_{G} is increasing. In the proof of Theorem C.1, the schedule was geometric in the parameter $u = e^t - 1$; here, the potential F_{G} replaces $\log u$ as the clock because $F_{\mathsf{G}}(v) - F_{\mathsf{G}}(u) = \log(v/u) + 3/u - 3/v = \log(1 + c_u(v))$. In particular, every step satisfies

$$c_{u_j}(u_{j+1}) = \frac{u_{j+1}}{u_j} e^{3/u_j - 3/u_{j+1}} - 1 \leqslant a_{\mathsf{G}}\,,$$

and therefore

$$\sum_{j=0}^{N_{\mathsf{G}}-1} c_{u_j}(u_{j+1})\big(\mathsf{DTC}(\pi_{u_j}) - \mathsf{DTC}(\pi_{u_{j+1}})\big) \leqslant a_{\mathsf{G}}\,\mathsf{DTC}(\pi_{u_0}) \leqslant a_{\mathsf{G}}\,\mathsf{DTC}(q) \leqslant \varepsilon/4\,,$$

since $\mathsf{DTC}(\pi_u)$ is nonincreasing in u by Lemma 2.4.

Next, we handle the contribution from score estimation error. We have

$$\log(1 + c_{u_j}(u_{j+1})) = \log(u_{j+1}/u_j) + 3/u_j - 3/u_{j+1} \leqslant \log(1 + a_{\mathsf{G}})\,,$$

so $c_{u_j}(u_{j+1}) \leqslant 1$ and thus by Eq. (19),

$$w_{u_j}(u_{j+1}) \leqslant 8\log(1 + c_{u_j}(u_{j+1}))\,.$$

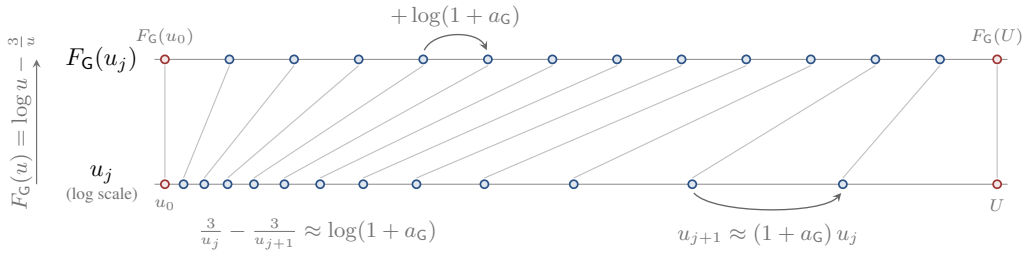


Figure 6: Step schedule for Gaussian diffusion sampler

Furthermore,

$$\sum_{j=0}^{N_{\mathsf{G}}-1} \log(1 + c_{u_j}(u_{j+1})) = F_{\mathsf{G}}(U) - F_{\mathsf{G}}(u_0) \leqslant \log(U/u_0) + 3/u_0 \leqslant \log(8L_{\mathsf{G}}U) + 24L_{\mathsf{G}} \lesssim L_{\mathsf{G}}\,. \tag{21}$$

Therefore,

$$\sum_{j=0}^{N_{\mathsf{G}}-1} w_{u_j}(u_{j+1})\, \varepsilon_{\mathsf{gauss}}(t(u_{j+1}))^2 \lesssim \varepsilon_{\mathsf{gauss}}^2 L_{\mathsf{G}}\,.$$

Next, we handle the initialization error. By Lemma D.10 and the choice of U,

$$\mathsf{KL}(\pi_U \parallel N(0, U\mathrm{Id}_{Sd})) \leqslant \frac{d}{2U} \leqslant \frac{\varepsilon}{8}\,.$$

Next, for the rounding error, we wish to apply Lemma D.11. By standard Gaussian tail bounds, $p^*(u_0) \leqslant Se^{-2L_{\mathsf{G}}} \leqslant 1/2$. Furthermore, $h_2(p^*(u)) \leqslant p^*(u)\log(e/p^*(u))$. Hence,

$$\mathbb{E}_{Z_{u_0}\sim\pi_{u_0}} \mathsf{KL}(\mathrm{law}(A \mid Z_{u_0}) \parallel Q_{u_0}(\cdot \mid Z_{u_0})) \leqslant dSe^{-2L_{\mathsf{G}}}(1 + 2L_{\mathsf{G}}) \leqslant dSe^{-L_{\mathsf{G}}} \leqslant \frac{\varepsilon}{4}\,.$$

Finally, we bound the total number of score oracle queries. By design, we have

$$N_{\mathsf{G}} \leq 1 + \frac{F_{\mathsf{G}}(U) - F_{\mathsf{G}}(u_0)}{\log(1 + a_{\mathsf{G}})}\,.$$

Recall from Eq. (21) that $F_{\mathsf{G}}(U) - F_{\mathsf{G}}(u_0) \lesssim L_{\mathsf{G}}$. Additionally, $1/\log(1 + a_{\mathsf{G}}) \leq 2/a_{\mathsf{G}} \lesssim 1 + \overline{\mathsf{DTC}}/\varepsilon$. Each application of $\widehat{R}^{\mathsf{G}}_{u_{j+1}\to u_j}$ requires only the score vector $\widehat{g}_{u_{j+1}}(y) \in \mathbb{R}^{Sd}$ at the current iterate y, from which Eq. (17) and the approximate reverse kernel $\widehat{R}^{\mathsf{G}}$ are formed. So the sampler makes N_{G} queries in total, which proves the query bound.

Substituting all of the above bounds into Lemma D.9 gives the claimed KL bound. □

E Construction of random empirical measure

In the second part of this paper, we turn to proving upper and lower bounds to separate the three paradigms of diffusion language modeling. Here, we first briefly present the family of distributions that we consider, which are based on a construction from recent work Xun & Price (2026) proving a query complexity lower bound for Gaussian diffusion.

Fix constants $0 < \kappa_- < \kappa_+ < \log 2$. In the subsequent sections, we will specialize to the case of binary alphabet, so that

$$\Sigma = \{\pm 1\}\,, \qquad \mathcal{X} = \{\pm 1\}^d\,, \qquad \beta_t = \frac{1 - e^{-t}}{2}\,.$$

Let $\mathcal{C} \subseteq \mathcal{X}$ denote the subset of the hypercube defining the codebook, and let $M = |\mathcal{C}|$ denote the size of the codebook. Throughout, we will take

$$M = \lceil e^{\kappa d} \rceil$$

for constant $\kappa \in [\kappa_-, \kappa_+]$. The compact interval $[\kappa_-, \kappa_+]$ ensures that all the critical values below remain in a fixed compact set.

Definition E.1. *Given a subset $\mathcal{C} \subseteq \{\pm 1\}^d$, denote the uniform distribution over $\mathcal{C}$ by*

$$q^{\mathcal{C}} := \mathrm{Unif}(\mathcal{C}) .$$

The random empirical measure *we consider is $q^{\mathcal{C}}$ for $\mathcal{C}$ a uniformly random M-element subset of $\{\pm 1\}^d$.*

We will often refer to the elements of $\mathcal{C}$ as codewords.

In all of our bounds in the sequel, we assume that the sampler:

- Knows the interval $[\kappa_-, \kappa_+]$ but does not know κ
- Knows that it has approximate score oracle access to a random empirical measure in the sense of Definition E.1, but does not know the specific realization of $\mathcal{C}$

Notation for subsequent sections. In our analysis, constants denoted by c, C may depend on the known parameters κ_-, κ_+, but on no other parameter; dependence on additional parameters is indicated by a subscript. For the uniform and Gaussian forward processes, we will overload the notation $q_t^{\mathcal{C}}$ to denote the law of X_t under either forward process, with $X_0 \sim q^{\mathcal{C}}$.

All three main results below as well as the uniform lower bound in Appendix I use Definition E.1. The upper bounds we prove for uniform and Gaussian diffusion hold for any $\kappa \in [\kappa_-, \kappa_+]$, whereas the lower bounds for masked and uniform diffusion hold even if κ is additionally known to be restricted to the lattice $\{(k/d)\log 2 : k \in \mathbb{Z}\}$, in which case $M = e^{\kappa d} = 2^k$.

Define the scalar mutual information for the uniform and Gaussian channels by

$$\mathsf{I}_{\mathsf{unif}}(t) := \frac{1+e^{-t}}{2}\log(1+e^{-t}) + \frac{1-e^{-t}}{2}\log(1-e^{-t}),$$
$$\mathsf{I}_{\mathsf{gauss}}(t) := I\left(V; e^{-t}V + \sqrt{1-e^{-2t}}\,G\right), \qquad V \sim \mathrm{Unif}(\{\pm 1\}), \quad G \sim N(0,1) .$$

Here and below, $h_2(p) := -p\log p - (1-p)\log(1-p)$ denotes the binary entropy function (in nats); note that $\mathsf{I}_{\mathsf{unif}}(t) = \log 2 - h_2(\beta_t)$.

Define critical times $t^\star_{\mathsf{unif}}$ and $t^\star_{\mathsf{gauss}}$ to be the unique solutions of

$$d\,\mathsf{I}_{\mathsf{unif}}(t^\star_{\mathsf{unif}}) = \log M \qquad \text{and} \qquad d\,\mathsf{I}_{\mathsf{gauss}}(t^\star_{\mathsf{gauss}}) = \log M . \tag{22}$$

Additionally, let $0 < \tau_- < \tau_+$ be (known) constants such that $t^\star_{\mathsf{unif}}, t^\star_{\mathsf{gauss}} \in [\tau_-, \tau_+]$ for every $\kappa \in [\kappa_-, \kappa_+]$ and all sufficiently large d, and define

$$I_0 = [\tau_-/2, \tau_+ + 1] . \tag{23}$$

In the proofs of our *upper* bounds, all times queried by our algorithms will lie in I_0 defined in Eq. (23), though we place no such constraint in the proofs of our *lower* bounds.

F $\widetilde{O}(\sqrt{d})$ UPPER BOUND FOR UNIFORM DIFFUSION

In this section, we give an algorithm for approximately sampling from random empirical measures using $\widetilde{O}(\sqrt{d})$ queries to any approximate uniform diffusion score oracle for such a distribution. This query complexity is tight: in Appendix I we prove a matching lower bound.

Theorem F.1 (Uniform diffusion with approximate scores)**.** *There are constants $B, C, c_0, d_0 > 0$, depending only on $[\kappa_-, \kappa_+]$, such that the following holds. Let $d \geq d_0$, let $\varepsilon, \delta \in (0, 1/4)$, and set*

$$\Lambda = \log \frac{Cd}{\varepsilon\delta} . \tag{24}$$

Suppose $\Lambda \leqslant c_0 d$.

Given query access to any approximate uniform diffusion score oracle $(\widehat{s}_t)$ for $q^{\mathcal{C}}$ with score error satisfying

$$\sup_{t\in I_0} \varepsilon_{\mathsf{unif}}(t) \leqslant e^{-B\Lambda} = \left(\frac{\varepsilon\delta}{Cd}\right)^B , \tag{25}$$

there is an algorithm using at most $O(\sqrt{d\Lambda}/\varepsilon^2)$ score oracle queries such that with probability at least $1-\delta$ over $\mathcal{C}$ which is sampled uniformly from all M-element subsets of $\mathcal{X}$, the output law of the algorithm $\widehat{q}^{\mathcal{C}}$ satisfies $\mathsf{TV}(\widehat{q}^{\mathcal{C}}, q^{\mathcal{C}}) \leqslant \varepsilon$.

In particular, for constant ε and any polynomially small $\delta = d^{-\Theta(1)}$, polynomially small score accuracy suffices, and the algorithm uses $O(\sqrt{d\log d})$ score queries.

For the rest of the section, Λ has the value in Eq. (24). We may decrease c_0 and increase d_0 later when necessary so that $\log d \leqslant \Lambda \leqslant c_0 d$.

In Section F.1, we quantify the critical window: above the window, $q_t^{\mathcal{C}}$ is close to the uniform distribution, while below the window the original sample can be recovered from its noisy observation. In Section F.2, we show how to locate this window using approximate score oracle queries at points sampled from $\mathrm{Unif}(\mathcal{X})$. Finally, in Section F.3, we describe how to simulate the reverse process across the window and then decode the codeword, completing the proof of Theorem F.1.

F.1 Quantifying the critical window

In this section, we characterize the critical window of noise levels such that, for higher noise levels t, the score of $q_t^{\mathcal{C}}$ is close to that of the uniform distribution, and for lower noise levels t, given a noisy sample from $q_t^{\mathcal{C}}$, it is possible to decode the original sample from $q^{\mathcal{C}}$ that gave rise to it with high accuracy.

Throughout, we will denote the likelihood ratio between $q_t^{\mathcal{C}}$ and the uniform distribution over $\mathcal{X}$ by

$$L_t(\cdot) \coloneqq \frac{q_t^{\mathcal{C}}(\cdot)}{2^{-d}} .$$

First, we bound the expected distance between $q_t^{\mathcal{C}}$ and the uniform distribution for large t.

Lemma F.2. *For any integer radius $0 \leqslant r \leqslant d$,*

$$\mathbb{E}_{\mathcal{C}}\, \mathsf{TV}(q_t^{\mathcal{C}}, \mathrm{Unif}(\mathcal{X})) \leqslant \frac{1}{2\sqrt{M}}(1+e^{-t})^{d/2}\cdot \tanh(t/2)^{(r+1)/2} + \mathbb{P}[\mathrm{Bin}(d,\beta_t) \leqslant r] .$$

We defer the proof to Appendix J.

Next, we show that given a noisy sample from $q_t^{\mathcal{C}}$ for small t, it is possible to recover the original sample from $q^{\mathcal{C}}$ with high accuracy.

Lemma F.3. *Given $\mathcal{C} \subseteq \{\pm 1\}^d$, define the Bayes-optimal recovery error*

$$\mathrm{err}_{\mathcal{C}}^{\mathsf{U}}(t) \coloneqq \inf_{\widehat{z}(\cdot)} \mathop{\mathbb{P}}_{z\sim q^{\mathcal{C}}, x\sim (K_t^{\mathsf{U}})^{\otimes d}(\cdot|z)} [\widehat{z}(x) \neq z] .$$

For any integer radius $0 \leqslant r \leqslant d$,

$$\mathbb{E}_{\mathcal{C}}\, \mathrm{err}_{\mathcal{C}}^{\mathsf{U}}(t) \leqslant M\cdot \mathbb{P}[\mathrm{Bin}(d,1/2) \leqslant r] + \mathbb{P}[\mathrm{Bin}(d,\beta_t) > r] .$$

We defer the proof to Appendix J.

Setting parameters appropriately and using the two estimates above, we can thus characterize the critical window for sampling from the random empirical measure using uniform diffusion: above the critical window, the forward process has sufficiently mixed so that the law of X_t is close to uniform, while below the critical window, the Bayes-optimal decoder can recover the original codeword with high probability.

Proposition F.4 (Critical window). *For every $K > 0$ there is $A_K > 0$ with the following property. For every $A \geq A_K$, there are $c_{K,A}, d_{K,A} > 0$, depending only on A, K, and $[\kappa_-, \kappa_+]$, such that the following holds whenever $d \geq d_{K,A}$ and $\log d \leqslant \Lambda \leqslant c_{K,A} d$. Set*

$$w = A\sqrt{\frac{\Lambda}{d}}\,. \tag{26}$$

Then we have

$$\begin{aligned} \mathbb{E}_{\mathcal{C}} \mathsf{TV}(q_t^{\mathcal{C}}, \mathrm{Unif}(\mathcal{X})) &\leqslant e^{-K\Lambda}, & t &\geq t^\star_{\mathsf{unif}} + w, \\ \mathbb{E}_{\mathcal{C}} \mathrm{err}_{\mathcal{C}}^{\mathsf{U}}(t) &\leqslant e^{-K\Lambda}, & 0 &\leqslant t \leqslant t^\star_{\mathsf{unif}} - w\,. \end{aligned}$$

Proof. We first observe that $\mathrm{err}_{\mathcal{C}}^{\mathsf{U}}(t)$ is nondecreasing in t and, by data processing inequality, $\mathsf{TV}(q_t^{\mathcal{C}}, \mathrm{Unif}(\mathcal{X}))$ is nonincreasing in t. Therefore, it suffices to consider the endpoints $t = t^\star_{\mathsf{unif}} + w$ and $t = t^\star_{\mathsf{unif}} - w$. By decreasing $c_{K,A}$ we may assume $w \leqslant \min\{\tau_-/2, 1\}$; then these points are within I_0, so β_t is bounded away from 0 and $1/2$ and

$$cw \leqslant |\beta_t - \beta_{t^\star_{\mathsf{unif}}}| \leqslant Cw, \qquad \text{for } t \in \{t^\star_{\mathsf{unif}} - w, t^\star_{\mathsf{unif}} + w\}\,, \tag{27}$$

because $\frac{\mathrm{d}}{\mathrm{d}t}\beta_t = e^{-t}/2$ is bounded above and below by positive constants on I_0. Here $c, C > 0$ depend only on $[\kappa_-, \kappa_+]$.

Above the critical window. Set $t = t^\star_{\mathsf{unif}} + w$, $\Delta = \beta_t - \beta_{t^\star_{\mathsf{unif}}}$, and truncation radius $r = \lfloor d(\beta_t + \beta_{t^\star_{\mathsf{unif}}})/2 \rfloor$. The radius is at least $d\Delta/2$ below the mean $d\beta_t$ of $\mathrm{Bin}(d, \beta_t)$. Hoeffding's inequality therefore gives

$$\mathbb{P}[\mathrm{Bin}(d, \beta_t) \leqslant r] \leqslant e^{-d\Delta^2/2}\,. \tag{28}$$

It remains to bound the first term in Lemma F.2. Eq. (22) gives $\log M = d(\log 2 - h_2(\beta_{t^\star_{\mathsf{unif}}}))$. Rewrite $1 + e^{-t} = 2(1 - \beta_t)$ and $\tanh(t/2) = \beta_t/(1 - \beta_t)$. Since $r + 1 \geq d\beta_{t^\star_{\mathsf{unif}}} + d\Delta/2$,

$$\begin{aligned} \log\left(\frac{(1+e^{-t})^d \tanh(t/2)^{r+1}}{M}\right) &\leqslant d\log\big(2(1-\beta_t)\big) + d\left(\beta_{t^\star_{\mathsf{unif}}} + \frac{\Delta}{2}\right)\log\frac{\beta_t}{1-\beta_t} - \log M \\ &= -d\,\mathsf{KL}(\mathrm{Ber}(\beta_{t^\star_{\mathsf{unif}}}) \,\|\, \mathrm{Ber}(\beta_t)) - \frac{d\Delta}{2}\log\frac{1-\beta_t}{\beta_t} \\ &\leqslant -cd\Delta\,. \end{aligned} \tag{29}$$

Lemma F.2 and Eqs. (28)–(29) now give

$$\mathbb{E}_{\mathcal{C}} \mathsf{TV}(q_t^{\mathcal{C}}, \mathrm{Unif}(\mathcal{X})) \leqslant \tfrac{1}{2}e^{-cd\Delta} + e^{-d\Delta^2/2}\,.$$

Below the critical window. Set $t = t^\star_{\mathsf{unif}} - w$, $\Delta = \beta_{t^\star_{\mathsf{unif}}} - \beta_t$, and again $r = \lfloor d(\beta_t + \beta_{t^\star_{\mathsf{unif}}})/2 \rfloor$. Since $r + 1 > d\beta_t + d\Delta/2$, Hoeffding's inequality gives

$$\mathbb{P}[\mathrm{Bin}(d, \beta_t) > r] \leqslant e^{-d\Delta^2/2}\,. \tag{30}$$

To control the chance of another point entering the ball, a standard binomial tail bound gives

$$\begin{aligned} M\,\mathbb{P}[\mathrm{Bin}(d, 1/2) \leqslant r] &\leqslant \exp\big(\log M - d\big(\log 2 - h_2((\beta_t + \beta_{t^\star_{\mathsf{unif}}})/2)\big)\big) \\ &= \exp\big(-d\big(h_2(\beta_{t^\star_{\mathsf{unif}}}) - h_2((\beta_t + \beta_{t^\star_{\mathsf{unif}}})/2)\big)\big) \leqslant e^{-cd\Delta}\,. \end{aligned} \tag{31}$$

In the last inequality, we used the fact that $\beta_{t^\star_{\mathsf{unif}}} - (\beta_t + \beta_{t^\star_{\mathsf{unif}}})/2 = \Delta/2$, and that $h_2'(p) = \log((1-p)/p)$ is bounded away from 0 for p between $(\beta_t + \beta_{t^\star_{\mathsf{unif}}})/2$ and $\beta_{t^\star_{\mathsf{unif}}}$. Combining Eqs. (30) and (31) with Lemma F.3 yields

$$\mathbb{E}_{\mathcal{C}} \mathrm{err}_{\mathcal{C}}^{\mathsf{U}}(t) \leqslant e^{-cd\Delta} + e^{-d\Delta^2/2}\,.$$

Finally, Eq. (27) implies $d\Delta^2 \geq cA^2\Lambda$. Since $w \leqslant 1$, $d\Delta \geq cdw \geq cdw^2 = cA^2\Lambda$. Choosing A large enough in terms of K makes $e^{-cd\Delta} + e^{-d\Delta^2/2}$ at most $e^{-K\Lambda}$. The monotonicity of $\mathsf{TV}(q_t^{\mathcal{C}}, \mathrm{Unif}(\mathcal{X}))$ and $\mathrm{err}_{\mathcal{C}}^{\mathsf{U}}(t)$ in t then gives the claimed bounds. □

For the rest of the section, let $w = A\sqrt{\Lambda/d}$ as in Eq. (26), where $A > 0$ is a constant to be fixed below. We may decrease c_0 and increase d_0 later when necessary.

F.2 IDENTIFYING THE CRITICAL WINDOW

Having characterized the critical window, we now give an algorithm for detecting its location given approximate uniform diffusion score access to $q^{\mathcal{C}}$.

Given $x \in \mathcal{X}$, define the denoiser $m_t(x) \coloneqq \mathbb{E}[X_0 \mid X_t = x]$. We will consider the following statistic:

$$T_t(x) \coloneqq \langle x, m_t(x)\rangle - de^{-t} = \frac{1-e^{-2t}}{2e^{-t}} \sum_{i=1}^{d} (1 - s_t(x)[i, -x_i])\,.$$

Intuitively, $\langle x, m_t(x)\rangle$ corresponds to the overlap between x and its denoising with respect to the true data distribution $q^{\mathcal{C}}$, and de^{-t} is the same but under denoising with respect to the null distribution $\mathrm{Unif}(\mathcal{X})$. As t goes to ∞, the difference between these tends to zero, but as t goes to 0, if x is a uniformly random point from $\mathcal{X}$, then we expect the denoiser under $q^{\mathcal{C}}$ to achieve worse overlap than the denoiser under $\mathrm{Unif}(\mathcal{X})$. This is formalized in the following:

Proposition F.5. *There are constants $a_0, b_0 > 0$, depending only on $[\kappa_-, \kappa_+]$, with $b_0 > 4a_0$, such that for every $K > 0$ there is $A_K > 0$ with the following property. For every $A \geq A_K$, there are $c_{K,A}, d_{K,A} > 0$ such that the following holds whenever $d \geq d_{K,A}$ and $\log d \leqslant \Lambda \leqslant c_{K,A} d$:*

$$\mathop{\mathbb{P}}_{\mathcal{C},x\sim\mathrm{Unif}(\mathcal{X})}[T_t(x) < -a_0 dw] \leqslant e^{-K\Lambda}\,, \qquad t \geq t^\star_{\mathsf{unif}} + 2w\,,$$
$$\mathop{\mathbb{P}}_{\mathcal{C},x\sim\mathrm{Unif}(\mathcal{X})}[T_t(x) > -b_0 dw] \leqslant e^{-K\Lambda}\,, \qquad t \leqslant t^\star_{\mathsf{unif}} - 2w\,,$$

uniformly over $t \in I_0$.

Proof. The overlap $\langle x, m_t(x)\rangle$ can be expressed succinctly as $A'_x(\theta(t))$ where

$$A_x(\theta) \coloneqq \log \mathbb{E}_{z\sim\mathrm{Unif}(\mathcal{C})}[e^{\theta\langle x,z\rangle}]\,,$$

and $\theta(t) \coloneqq \mathrm{atanh}(e^{-t})$. Also define

$$B(\theta) \coloneqq d \log\cosh(\theta)\,,$$

observing that $B'(\theta(t)) = de^{-t}$. We have

$$\log L_t(x) = A_x(\theta(t)) - B(\theta(t)) \qquad \text{and} \qquad T_t(x) = A'_x(\theta(t)) - B'(\theta(t))\,.$$

High-noise regime ($t \geq t^\star_{\mathsf{unif}} + 2w$). Choose neighboring noise levels t and $t^+ = t + \eta w$ and define $\theta = \theta(t)$ and $\theta^+ = \theta(t^+)$ so that $0 \leqslant \theta - \theta^+ \asymp \eta w$. The function $A_x(\theta)$ is convex in θ, so

$$A'_x(\theta) \geq \frac{A_x(\theta) - A_x(\theta^+)}{\theta - \theta^+}\,,$$

and thus

$$T_t(x) \geq \frac{\log L_t(x) - \log L_{t^+}(x)}{\theta - \theta^+} + \Big(\frac{B(\theta) - B(\theta^+)}{\theta - \theta^+} - B'(\theta)\Big)\,.$$

Because $B''(u) = d\,\mathrm{sech}^2(u) \leqslant d$, by Taylor's theorem the second term above is $\geq -C' d\eta w$ for some constant $C' > 0$. On the other hand, since $\mathbb{E}_{X\sim\mathrm{Unif}(\mathcal{X})}|L_t(X) - 1| = 2\mathsf{TV}(q^{\mathcal{C}}_t, \mathrm{Unif}(\mathcal{X}))$, Proposition F.4 and Markov's inequality show that $L_t(X), L_{t^+}(X) \in [1/2, 3/2]$ except with probability at most $e^{-(K+1)\Lambda}$. On this event, the first term is $\geq -C''/(\eta w)$ for some constant $C'' > 0$. With constant η, $1/(\eta w) = o(dw)$ since $dw^2 = A^2\Lambda = \Omega(\log d)$. Choosing constant η small enough, we can ensure $-C''/(\eta w) - C' d\eta w \geq -a_0 dw$ for any prescribed constant $a_0 > 0$.

Low-noise regime ($t \leqslant t^\star_{\mathsf{unif}} - 2w$). First note that $A'_x(\theta(t))$ is an average of overlaps of the form $\langle x, z\rangle$ for $z \in \mathcal{C}$. We bound the max over these overlaps by a Chernoff and union bound calculation. By Chernoff, for any $u > 0$,

$$\mathop{\mathbb{P}}_{z\sim\mathrm{Unif}(\mathcal{X})}[\langle x, z\rangle \geq du] \leqslant e^{-d\,\mathsf{KL}(\mathrm{Ber}(\frac{1+u}{2}) \,\|\, \mathrm{Ber}(1/2))}\,.$$

So for $u = e^{-t^\star_{\mathsf{unif}}} + O(\Lambda/d)$, noting that $\mathsf{KL}(\mathrm{Ber}(\frac{1+u}{2}) \parallel \mathrm{Ber}(1/2)) = \mathsf{I}_{\mathsf{unif}}(\log(1/u))$ and recalling that $\mathsf{I}_{\mathsf{unif}}(t^\star_{\mathsf{unif}}) = \log(M)/d$, we conclude by a union bound over $\mathcal{C}$ that with probability at least $1 - e^{-\Theta(\Lambda)}$ over the randomness of $\mathcal{C}$,

$$\max_{z \in \mathcal{C}} \langle x, z \rangle \leqslant d e^{-t^\star_{\mathsf{unif}}} + O(\Lambda)\,.$$

Henceforth condition on this event. Then

$$T_t(x) \leqslant d e^{-t^\star_{\mathsf{unif}}} - d e^{-t} + O(\Lambda)\,.$$

When $t \leqslant t^\star_{\mathsf{unif}} - 2w$, we have $d e^{-t^\star_{\mathsf{unif}}} - d e^{-t} \leqslant -b' dw$ for some constant $b' > 0$. On the other hand, since $\Lambda \leqslant c_{K,A} d$ gives $\Lambda \leqslant (\sqrt{c_{K,A}}/A)\, dw$, the $O(\Lambda)$ term is at most $b'' dw$ for some constant $b'' \in (0, b')$ once $c_{K,A}$ is small enough. Taking $b_0 = b' - b''$, the second claimed bound follows. Take a_0 small enough so that $b_0 > 4a_0$. □

The quantity $T_t(x)$ is therefore a good statistic for identifying the critical window. In practice, we will not have access to $m_t(x)$ or $s_t(x)$, but we can estimate $T_t(x)$ using estimated scores $\widehat{s}_t(x)$, which we will show are accurate enough to identify the critical window.

We first define the empirical version of $T_t(x)$. Let $\ell_t = \beta_t/(1-\beta_t)$. For a fixed initial point z, flipping one coordinate of x changes $(K_t^{\mathsf{U}})^{\otimes d}(x \mid z)$ by a factor of either ℓ_t or ℓ_t^{-1}. The score is a weighted average of these ratios, with weights given by the posterior distribution of z. So for every $t > 0$, $x \in \mathcal{X}$, and $i \in [d]$,

$$\ell_t \leqslant s_t(x)[i, -x_i] \leqslant \ell_t^{-1}\,. \tag{32}$$

In place of the Bregman projection from Section C, here we simply *clip* the entries of the score vector to this range to get the *clipped approximate score*

$$\widetilde{s}_t(x)[i, -x_i] \coloneqq \min\big(\ell_t^{-1}, \max(\ell_t, \widehat{s}_t(x)[i, -x_i])\big)\,. \tag{33}$$

Define the following approximation to T_t that the sampler can actually form using approximate score access:

$$\widehat{T}_t(x) \coloneqq \frac{1 - e^{-2t}}{2e^{-t}} \sum_{i=1}^{d} \big(1 - \widetilde{s}_t(x)[i, -x_i]\big)\,. \tag{34}$$

The statistic $\widehat{T}_t(x)$ can be computed with only one approximate score oracle query at x. Here we show that $\widehat{T}_t(x)$ is a good estimate of $T_t(x)$. We will quantify this relative to the squared Euclidean error of the clipped score:

$$F_t(x) \coloneqq \sum_{i=1}^{d} \big(\widetilde{s}_t(x)[i, -x_i] - s_t(x)[i, -x_i]\big)^2\,. \tag{35}$$

Lemma F.6. *Uniformly over $t \in I_0$ and $x \in \mathcal{X}$, we have*

$$|\widehat{T}_t(x) - T_t(x)| \leqslant C\sqrt{d F_t(x)}\,, \qquad \mathbb{E}_{X \sim q_t^{\mathcal{C}}} F_t(X) \leqslant C' \varepsilon_{\mathsf{unif}}(t) \tag{36}$$

for constants $C, C' > 0$ depending only on $[\kappa_-, \kappa_+]$.

We defer the proof to Appendix J.

We record the following standard estimates for Hamming-sphere sizes and binomial point masses, which will be used repeatedly below.

Lemma F.7. *For $1 \leqslant r \leqslant d-1$ and $v = r/d$,*

$$\frac{e^{d h_2(v)}}{d+1} \leqslant \binom{d}{r} \leqslant e^{d h_2(v)}\,. \tag{37}$$

Consequently, for any $p \in (0,1)$,

$$\mathbb{P}[\mathrm{Bin}(d,p) = r] \geq \frac{1}{d+1} \exp\left(-d\,\mathsf{KL}(\mathrm{Ber}(v) \parallel \mathrm{Ber}(p))\right) \geq \frac{1}{d+1} \exp\left(-\frac{d(v-p)^2}{p(1-p)}\right)\,. \tag{38}$$

Moreover, for all $v, p \in (0,1)$,

$$\mathsf{KL}(\mathrm{Ber}(v) \parallel \mathrm{Ber}(p)) \leqslant \frac{(v-p)^2}{p(1-p)}\,. \tag{39}$$

We defer the proof to Appendix J.

The following lemma allows us to control the likelihood ratio $L_t(x)$ near the critical time $t^\star_{\mathsf{unif}}$.

Lemma F.8. *For every $A, K, L > 0$, there are constants $C_{A,K,L}, c_{A,K,L}, d_{A,K,L} > 0$, depending only on A, K, L and $[\kappa_-, \kappa_+]$, such that whenever $d \geq d_{A,K,L}$, $\log d \leqslant \Lambda \leqslant c_{A,K,L} d$, and $|t - t^\star_{\mathsf{unif}}| \leqslant Lw$, we have*

$$\mathop{\mathbb{P}}_{\mathcal{C}}[L_t(x) < e^{-C_{A,K,L}\Lambda}] \leqslant e^{-K\Lambda}\,,$$

uniformly over $x \in \mathcal{X}$.

We defer the proof to Appendix J.

We can now combine all of the ingredients above to prove that $\widehat{T}_t(x)$ can be used to identify the critical window.

Proposition F.9. *There are constants $0 < a < b$, depending only on $[\kappa_-, \kappa_+]$, such that for every $K > 0$ there is $A_K > 0$ with the following property. For every $A \geq A_K$, there are $B_{K,A}, c_{K,A}, d_{K,A} > 0$ such that the following holds whenever $d \geq d_{K,A}$, $\log d \leqslant \Lambda \leqslant c_{K,A} d$, $B \geq B_{K,A}$, and the score oracle satisfies Eq. (25):*

$$\mathop{\mathbb{P}}_{\mathcal{C}, x\sim\mathrm{Unif}(\mathcal{X})}[\widehat{T}_t(x) < -adw] \leqslant e^{-K\Lambda}\,, \qquad t \geq t^\star_{\mathsf{unif}} + 2w\,, \tag{40}$$

$$\mathop{\mathbb{P}}_{\mathcal{C}, x\sim\mathrm{Unif}(\mathcal{X})}[\widehat{T}_t(x) > -bdw] \leqslant e^{-K\Lambda}\,, \qquad t^\star_{\mathsf{unif}} - 3w \leqslant t \leqslant t^\star_{\mathsf{unif}} - 2w\,. \tag{41}$$

uniformly over $t \in I_0$.

Proof. Let a_0, b_0 be the constants in Proposition F.5, chosen so that $b_0 > 4a_0$, and set

$$a := 2a_0 \qquad \text{and} \qquad b := b_0 - a_0\,.$$

Then $0 < a < b$.

By Lemma F.6, the event $|\widehat{T}_t - T_t| > a_0 dw$ implies $F_t > cdw^2$, and its probability under the true noised law $q_t^{\mathcal{C}}$ is at most $Ce^{-B\Lambda}/(dw^2)$.

High-noise regime ($t \geq t^\star_{\mathsf{unif}} + 2w$). Replacing $q_t^{\mathcal{C}}$ by the uniform law $\mathrm{Unif}(\mathcal{X})$ adds at most $\mathsf{TV}(q_t^{\mathcal{C}}, \mathrm{Unif}(\mathcal{X}))$. Averaging over $\mathcal{C}$ and using Proposition F.4 with exponent $K+2$ therefore gives an exponentially small error probability for a uniform query as well. Therefore, Markov's inequality implies

$$\mathop{\mathbb{P}}_{\mathcal{C}, x\sim\mathrm{Unif}(\mathcal{X})}[|\widehat{T}_t(x) - T_t(x)| > a_0 dw] \leqslant \frac{Ce^{-B\Lambda}}{dw^2} + e^{-(K+2)\Lambda}\,. \tag{42}$$

Since $dw^2 = A^2\Lambda \geq 1$, choosing B sufficiently large makes the right-hand side at most $e^{-(K+1)\Lambda}$ for all sufficiently large d.

On the other hand, Proposition F.5 with exponent $K+2$ gives

$$\mathop{\mathbb{P}}_{\mathcal{C}, x\sim\mathrm{Unif}(\mathcal{X})}[T_t(x) < -a_0 dw] \leqslant e^{-(K+2)\Lambda}\,. \tag{43}$$

If neither event in Eqs. (42) and (43) occurs, then

$$\widehat{T}_t(x) \geq T_t(x) - a_0 dw \geq -2a_0 dw = -adw\,.$$

Consequently,

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathcal{C}, x\sim\mathrm{Unif}(\mathcal{X})}[\widehat{T}_t(x) < -adw] \\
&\leqslant \mathop{\mathbb{P}}_{\mathcal{C}, x\sim\mathrm{Unif}(\mathcal{X})}[T_t(x) < -a_0 dw] + \mathop{\mathbb{P}}_{\mathcal{C}, x\sim\mathrm{Unif}(\mathcal{X})}[|\widehat{T}_t(x) - T_t(x)| > a_0 dw] \\
&\leqslant e^{-(K+2)\Lambda} + e^{-(K+1)\Lambda} \leqslant e^{-K\Lambda}
\end{aligned}$$

for all sufficiently large d, proving Eq. (40).

Low-noise regime ($t \in [t^\star_{\mathsf{unif}} - 3w, t^\star_{\mathsf{unif}} - 2w]$). Lemma F.8 (with $L = 3$) implies $L_t(x) \geq e^{-C\Lambda}$, where $C = C_{A,K',3}$, with probability at least $1 - e^{-K'\Lambda}$, where K' can be chosen arbitrarily large. On the event $L_t(x) \geq e^{-C\Lambda}$, we have the pointwise bound $2^{-d}\mathbb{1}[L_t(x) \geq e^{-C\Lambda}] \leqslant e^{C\Lambda} q_t^{\mathcal{C}}(x)$. Thus Markov's inequality gives

$$\mathbb{P}_{\mathcal{C},x\sim\mathrm{Unif}(\mathcal{X})}[|\widehat{T}_t(x) - T_t(x)| > a_0 dw] \leqslant e^{-K'\Lambda} + \frac{C' e^{-(B-C)\Lambda}}{dw^2}. \tag{44}$$

Choose $K' > K + 2$ and then with A fixed, choose B large enough so that the last display is at most $e^{-(K+1)\Lambda}$.

Proposition F.5 with exponent $K + 2$ gives that throughout this interval

$$\mathbb{P}_{\mathcal{C},x\sim\mathrm{Unif}(\mathcal{X})}[T_t(x) > -b_0 dw] \leqslant e^{-(K+2)\Lambda}, \tag{45}$$

since every such t satisfies $t \leqslant t^\star_{\mathsf{unif}} - 2w$. If neither the event in Eq. (44) nor the event in Eq. (45) occurs, then

$$\widehat{T}_t(x) \leqslant T_t(x) + a_0 dw \leqslant -(b_0 - a_0)dw = -bdw.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}_{\mathcal{C},x\sim\mathrm{Unif}(\mathcal{X})}[\widehat{T}_t(x) > -bdw] \\
&\leqslant \mathbb{P}_{\mathcal{C},x\sim\mathrm{Unif}(\mathcal{X})}[T_t(x) > -b_0 dw] + \mathbb{P}_{\mathcal{C},x\sim\mathrm{Unif}(\mathcal{X})}[|\widehat{T}_t(x) - T_t(x)| > a_0 dw] \\
&\leqslant e^{-(K+2)\Lambda} + e^{-(K+1)\Lambda} \leqslant e^{-K\Lambda}
\end{aligned}$$

for all sufficiently large d. This proves Eq. (41). □

From now on, fix $A \geq A_{12}$ and choose $B \geq B_{12,A}$. We may decrease c_0 and increase d_0 so that $c_0 \leqslant c_{12,A}$, $d_0 \geq d_{12,A}$, and $w \leqslant \min(\tau_-/16, 1/16)$. We further decrease c_0 and increase d_0 below whenever necessary.

Algorithm 1: Finding the uniform diffusion critical window

Input: d, $[\kappa_-, \kappa_+]$, w, Λ, and query access to the score oracle $\widehat{s}$.
Output: An estimate $\widehat{t}$ of the critical time $t^\star_{\mathsf{unif}}$.

```
Let 𝒯 = {t_1 > t_2 > ··· > t_m} be a decreasing grid covering [τ_- − 3w, τ_+ + 3w] with mesh w/20;
Set R ← 2⌈CΛ⌉ + 1;
for t ∈ 𝒯 do
    for j ← 1 to R do
        Sample X^(j) ~ Unif(𝒳);
        Query ŝ_t(X^(j)) and compute T̂_t(X^(j)) using Eqs. (33)–(34);
    if median{T̂_t(X^(1)), ..., T̂_t(X^(R))} < −((a+b)/2) dw then
        return t
return τ_-;
```

We can now conclude the proof of the main result of this subsection, namely an algorithm for estimating the critical noise level $t^\star_{\mathsf{unif}}$, essentially by scanning from high to low noise levels and stopping when the median of $\widehat{T}_t$ over enough random inputs becomes sufficiently negative (see Algorithm 1 for details).

Proposition F.10. *Under the hypotheses of Theorem F.1, Algorithm 1 uses $O(\Lambda/w)$ score oracle queries. For at least a $1 - \delta/4$ fraction of $\mathcal{C}$, it satisfies*

$$\mathbb{P}[t^\star_{\mathsf{unif}} - 3w \leqslant \widehat{t} \leqslant t^\star_{\mathsf{unif}} + 2w \mid \mathcal{C}] \geq 1 - \varepsilon/8. \tag{46}$$

We defer the proof to Appendix J.

F.3 Sampling after locating the window

Having found the window, we initialize right at the upper edge of the window, using the uniform distribution, and approximately simulate the reverse process until we reach its lower edge. Our algorithm for simulation is a naive discretization; unlike the DTC-adaptive sampler in Section C, we do not need to carefully tune the step sizes, and we use the clipped score in lieu of the Bregman projection.

First, we specialize the first part of Lemma B.1 to the binary alphabet setting that we consider here. For $0 < s < t$, write $h = t - s$ and define the affine function

$$p_h(r) = \frac{\beta_h}{\alpha_h}\big((1-\beta_h)r - \beta_h\big)\,. \tag{47}$$

Lemma B.1 implies that $p_h(s_t(y)[i, -y_i])$ is the probability that $(X_s)_i \neq y_i$ conditional on $X_t = y$. As in Section C, in a single step of our sampler it will flip coordinate i with probability

$$\widehat{p}^{\,i}_{t\to s}(y) = p_h\big(\widetilde{s}_t(y)[i, -y_i]\big)\,, \tag{48}$$

independently over i. The only distinctions from Section C are that here we use the clipped score $\widetilde{s}_t$, and because the alphabet is binary, we only have to specify a single probability for each coordinate. As before, we will denote the resulting approximate product reverse kernel by $\widehat{R}^{\mathsf{unif}}_{t\to s}$.

Next we bound the sampling error coming from using $\widetilde{s}_t$ instead of s_t in this product reverse kernel.

Lemma F.11. *There are constants $C, h_0 > 0$, depending only on $[\kappa_-, \kappa_+]$, such that whenever $s, t \in I_0$ and $0 < h = t - s \leqslant h_0$, the probabilities in Eq. (48) belong to $(0, 1)$. Denoting $P^i_{t\to s}(y, \cdot) = \mathrm{law}(X^i_s \mid X_t = y)$ and $\widehat{P}^i_{t\to s}(y, \cdot)$ the Bernoulli law defined by Eq. (48), we have*

$$\mathbb{E}_{Y\sim q^{\mathcal{C}}_t} \sum_{i=1}^{d} \mathsf{KL}(P^i_{t\to s}(Y) \parallel \widehat{P}^i_{t\to s}(Y)) \leqslant Ch\varepsilon_{\mathsf{unif}}(t)\,. \tag{49}$$

We defer the proof to Appendix J.

Finally, we control the overall error incurred by our sampler over the course of crossing the critical window. Here we re-use some of the basic chain rule calculations from Section C, but because we take steps of size $\asymp \varepsilon^2/d$, we do not need to tune the step sizes so as to take advantage of telescoping.

Proposition F.12. *Fix $t_- < t_+$ in I_0, let $\Delta = t_+ - t_-$, and let $r_j = t_- + j\Delta/N$ for $0 \leqslant j \leqslant N$. Suppose $h = \Delta/N \leqslant h_0$, and let $\widehat{R}$ be the composition of the approximate product reverse kernels from r_N down to r_0. If $\varepsilon_{\mathsf{unif}}(r_j) \leqslant \eta$ for $1 \leqslant j \leqslant N$, then any initial distribution ρ satisfies*

$$\mathsf{TV}(\rho\widehat{R}, q^{\mathcal{C}}_{t_-}) \leqslant \mathsf{TV}(\rho, q^{\mathcal{C}}_{t_+}) + C\sqrt{\frac{d\Delta}{N} + \Delta\eta}\,. \tag{50}$$

Proof. First initialize from $q^{\mathcal{C}}_{t_+}$. Let P be the true reverse path law on $(X_{r_N}, \ldots, X_{r_0})$ and let $\widehat{P}$ be the path law obtained from the approximate product kernels. Write $R_j(y, \cdot) = \mathrm{law}(X_{r_j} \mid X_{r_{j+1}} = y)$ and $\widehat{R}_j(y, \cdot) = \prod_i \widehat{P}^i_{r_{j+1}\to r_j}(y, \cdot)$. For every y, the identity

$$\mathsf{KL}(R_j(y,\cdot) \parallel \widehat{R}_j(y,\cdot)) = \mathsf{TC}(X_{r_j} \mid X_{r_{j+1}} = y) + \sum_{i=1}^{d} \mathsf{KL}(P^i_{r_{j+1}\to r_j}(y,\cdot) \parallel \widehat{P}^i_{r_{j+1}\to r_j}(y,\cdot))$$

separates discretization error and score approximation error. The KL chain rule, Lemma C.2 (with $s = r_j \geq t_-$), Lemma F.11, and telescoping therefore give

$$\begin{aligned}
\mathsf{KL}(P \parallel \widehat{P}) &\leqslant \frac{e^{2h}-1}{1-e^{-t_-}} \sum_{j=0}^{N-1} \big(\mathsf{DTC}(q^{\mathcal{C}}_{r_j}) - \mathsf{DTC}(q^{\mathcal{C}}_{r_{j+1}})\big) + \sum_{j=0}^{N-1} Ch\varepsilon_{\mathsf{unif}}(r_{j+1}) \\
&\leqslant C\frac{\Delta}{N}\big(\mathsf{DTC}(q^{\mathcal{C}}_{t_-}) - \mathsf{DTC}(q^{\mathcal{C}}_{t_+})\big) + C\frac{\Delta}{N}\sum_{j=0}^{N-1} \varepsilon_{\mathsf{unif}}(r_{j+1}) \\
&\leqslant C\frac{d\Delta}{N} + C\Delta\eta\,,
\end{aligned}$$

where $\mathsf{DTC}(q^{\mathcal{C}}_{t_-}) \leqslant H(q^{\mathcal{C}}_{t_-}) \leqslant d \log 2$. Data processing inequality and Pinsker's inequality imply Eq. (50) when the initial law is $q^{\mathcal{C}}_{t_+}$. Replacing it by ρ adds $\mathsf{TV}(\rho \widehat{R}, q^{\mathcal{C}}_{t_+} \widehat{R}) \leqslant \mathsf{TV}(\rho, q^{\mathcal{C}}_{t_+})$. □

Below the window, the posterior on the codeword that would have generated the noisy iterate is typically concentrated on a single point of $\mathcal{C}$. We can recover that point by taking the signs of the estimated posterior expectation. Use Lemma B.1 once more, now for the transition from t to 0, and define

$$\widehat{m}_t(y)_i = y_i \left[1 - 2p_t\big(\widetilde{s}_t(y)[i, -y_i]\big)\right], \qquad \widehat{z}_t(y) = \big(\mathrm{sgn}(\widehat{m}_t(y)_1), \dots, \mathrm{sgn}(\widehat{m}_t(y)_d)\big), \tag{51}$$

with p_t given by Eq. (47) with $h = t$.

Lemma F.13. *For $t \in I_0$, $X_0 \sim q^{\mathcal{C}}$, and its forward observation X_t, the rule in Eq. (51) satisfies*

$$\mathbb{P}[\widehat{z}_t(X_t) \neq X_0 \mid \mathcal{C}] \leqslant 5\mathrm{err}^{\mathsf{U}}_{\mathcal{C}}(t) + C\varepsilon_{\mathsf{unif}}(t)\,. \tag{52}$$

In particular, the law of $\widehat{z}_t(X_t)$ is within this distance of $q^{\mathcal{C}}$ in total variation.

Proof. Define a Bayes-optimal decoder (deterministic function) z_*: for each y, $z_*(y)$ is a point with largest posterior probability given $X_t = y$. Let $\pi_*(y)$ denote the corresponding posterior mass. The Bayes error is $\mathbb{E}[1 - \pi_*(X_t) \mid \mathcal{C}] = \mathrm{err}^{\mathsf{U}}_{\mathcal{C}}(t)$. Therefore by Markov's inequality,

$$\mathbb{P}[\pi_*(X_t) < 3/4 \mid \mathcal{C}] \leqslant 4\mathrm{err}^{\mathsf{U}}_{\mathcal{C}}(t)\,.$$

On the complementary event, every coordinate of the exact posterior expectation has the sign of $z_*(X_t)$ and magnitude at least $1/2$.

The map from a score to the corresponding posterior-mean coordinate in the first part of Eq. (51) is affine with bounded slope, since I_0 is compact. Lemma F.6 consequently gives

$$\mathbb{E}_{X_t \sim q^{\mathcal{C}}_t} \|\widehat{m}_t(X_t) - m_t(X_t)\|_2^2 \leqslant C\varepsilon_{\mathsf{unif}}(t)\,.$$

When $\pi_*(X_t) \geq 3/4$, a wrong sign requires the error in at least one coordinate to have magnitude at least $1/2$. Markov's inequality bounds the probability of this event by $C\varepsilon_{\mathsf{unif}}(t)$. Finally, $z_*(X_t) \neq X_0$ with probability $\mathrm{err}^{\mathsf{U}}_{\mathcal{C}}(t)$. These three possible failures give Eq. (52). The last assertion follows by coupling $\widehat{z}_t(X_t)$ with X_0. □

We are now ready to prove the main result of the section.

Algorithm 2: Sampling from the random empirical measure with uniform diffusion

Input: d, $[\kappa_-, \kappa_+]$, target accuracy ε, failure probability δ, and a score oracle satisfying Eq. (25).
Output: A sample from the approximate output law $\widehat{q}^{\mathcal{C}}$.

Set Λ as in Eq. (24) and w as in Eq. (26);
Run Algorithm 1 to obtain $\widehat{t}$;
Set $t_+ \leftarrow \widehat{t} + 4w$, $t_- \leftarrow \widehat{t} - 4w$, $N \leftarrow \lceil \frac{Cdw}{\varepsilon^2} \rceil$, $h \leftarrow \frac{8w}{N}$;
Set $r_j \leftarrow t_- + jh, \qquad 0 \leqslant j \leqslant N$;
Sample $Y_N \sim \mathrm{Unif}(\mathcal{X})$ independently of Algorithm 1;
for $j \leftarrow N$ **to** 1 **do**
    Query $\widehat{s}_{r_j}(Y_j)$ and compute $\widetilde{s}_{r_j}(Y_j)$ using Eq. (33);
    for $i \leftarrow 1$ **to** d **do**
        Set $\widehat{p}_{j,i} \leftarrow p_h\big(\widetilde{s}_{r_j}(Y_j)[i, -(Y_j)_i]\big)$;
        Independently draw $B_{j,i} \sim \mathrm{Ber}(\widehat{p}_{j,i})$ and set $(Y_{j-1})_i \leftarrow (-1)^{B_{j,i}}(Y_j)_i$;
Query $\widehat{s}_{t_-}(Y_0)$ and compute $\widehat{z}_{t_-}(Y_0)$ using Eqs. (33)–(51);
return $\widehat{z}_{t_-}(Y_0)$;

Proof of Theorem F.1. We analyze Algorithm 2. Proposition F.4, with exponent 12, and Markov's inequality imply

$$\mathsf{TV}(q^{\mathcal{C}}_{t_1}, \mathrm{Unif}(\mathcal{X})) \leqslant \varepsilon/16, \qquad \mathrm{err}^{\mathsf{U}}_{\mathcal{C}}(t_2) \leqslant \varepsilon/80\,,$$

uniformly over $t_1 \geq t^\star_{\mathsf{unif}} + w$ and $t_2 \leqslant t^\star_{\mathsf{unif}} - w$, outside an event over $\mathcal{C}$ of probability at most $\delta/2$. Proposition F.10 removes a further event over $\mathcal{C}$ of probability at most $\delta/4$ where the search of the critical window fails.

Now fix a $\mathcal{C}$ outside these failure events. Conditional on this $\mathcal{C}$, the window search succeeds with probability at least $1 - \varepsilon/8$.

Now suppose Algorithm 1 succeeds, and condition on its returned value $\widehat{t}$. The successful search ensures that

$$t_+ \geq t^\star_{\mathsf{unif}} + w, \qquad t_- \leqslant t^\star_{\mathsf{unif}} - w, \qquad t_+ - t_- = 8w\,.$$

Both t_+ and t_- belong to I_0 by the choice of w.

Now we proceed to analyze the law of Y_0 generated by applying approximate reverse kernels to the uniform sample Y_N in Algorithm 2. Since $N \geq Cdw/\varepsilon^2$ and $\varepsilon \leqslant 1/4$, we have $h = \frac{8w}{N} \leqslant \frac{8\varepsilon^2}{Cd} \leqslant \frac{1}{2Cd}$. Thus, by increasing d_0 if necessary, we have $h \leqslant h_0$, as required by Proposition F.12. Therefore, Proposition F.12 gives

$$\mathsf{TV}(\mathrm{law}(Y_0 \mid \mathcal{C}, \widehat{t}), q^{\mathcal{C}}_{t_-}) \leqslant \varepsilon/16 + C\sqrt{\frac{8dw}{N} + 8we^{-B\Lambda}} \leqslant \frac{3\varepsilon}{16}\,, \tag{53}$$

after choosing B and the constant in N sufficiently large.

Finally consider the final recovery step. If the input law were exactly $q^{\mathcal{C}}_{t_-}$, Lemma F.13 and Eq. (52) would bound the distance of its output law from $q^{\mathcal{C}}$ by $5\varepsilon/80 + Ce^{-B\Lambda} \leqslant \varepsilon/8$. Combining this with Eq. (53) (after applying the same recovery rule) gives

$$\mathsf{TV}(\mathrm{law}(\widehat{z}_{t_-}(Y_0) \mid \mathcal{C}), q^{\mathcal{C}}) \leqslant \frac{5\varepsilon}{16}\,,$$

uniformly over every possible successful value of $\widehat{t}$.

Algorithm 1 fails with probability at most $\varepsilon/8$. Thus $\mathsf{TV}(\widehat{q}^{\mathcal{C}}, q^{\mathcal{C}}) \leqslant 7\varepsilon/16 < \varepsilon$. The total failure probability over $\mathcal{C}$ is at most $3\delta/4 < \delta$.

Finally, the window search stage uses $O(\Lambda/w)$ queries, the reverse simulation stage uses $N = O(dw/\varepsilon^2 + 1)$, and the final recovery stage uses one. Substituting $w = A\sqrt{\Lambda/d}$ gives the claimed query complexity. □

G $\widetilde{O}(\sqrt{d})$ UPPER BOUND FOR GAUSSIAN DIFFUSION

In this section, we give an algorithm for approximately sampling from random empirical measures using $\widetilde{O}(\sqrt{d})$ queries to any sufficiently accurate Gaussian diffusion score oracle for such a distribution. This query complexity is also tight: a matching lower bound was proved in Xun & Price (2026).

Theorem G.1 (Gaussian diffusion with approximate scores)**.** *There are constants $B, C, c_0, d_0 > 0$, depending only on $[\kappa_-, \kappa_+]$, such that the following holds. Let $d \geq d_0$, let $\varepsilon, \delta \in (0, 1/4)$, and set*

$$\Lambda = \log \frac{Cd}{\varepsilon\delta}\,. \tag{54}$$

Suppose $\Lambda \leqslant c_0 d$.

Given query access to any approximate Gaussian diffusion score oracle $(\widehat{s}_t)$ for $q^{\mathcal{C}}$ with score error satisfying

$$\sup_{t \in I_0} \varepsilon_{\mathsf{gauss}}(t) \leqslant e^{-B\Lambda} = \left(\frac{\varepsilon\delta}{Cd}\right)^B\,, \tag{55}$$

there is an algorithm using at most $O(\sqrt{d\Lambda}/\varepsilon^2)$ score oracle queries such that with probability at least $1 - \delta$ over $\mathcal{C}$ sampled uniformly from all M-element subsets of $\mathcal{X}$, the output law of the algorithm $\widehat{q}^{\mathcal{C}}$ satisfies $\mathsf{TV}(\widehat{q}^{\mathcal{C}}, q^{\mathcal{C}}) \leqslant \varepsilon$.

In particular, for constant ε and any polynomially small $\delta = d^{-\Theta(1)}$, polynomially small score accuracy suffices, and the algorithm uses $O(\sqrt{d \log d})$ score queries.

For the rest of the section, Λ has the value in Eq. (54). We may decrease c_0 and increase d_0 later when necessary so that $\log d \leqslant \Lambda \leqslant c_0 d$.

Define the *reference distribution*

$$\nu_t := \mathrm{law}(e^{-t}U + \sigma_t G), \qquad U \sim \mathrm{Unif}(\mathcal{X}), \quad G \sim N(0, I_d)\,. \tag{56}$$

Thus ν_t is obtained by applying the same forward process to the uniform distribution on the whole cube. The algorithm can of course sample from ν_t without knowing $\mathcal{C}$.

In Section G.1, we quantify the critical window: above the window, $q_t^{\mathcal{C}}$ is close to the reference distribution ν_t, while below the window the original sample can be recovered from its noisy observation. In Section G.2, we show how to locate this window using approximate score oracle queries at points sampled from ν_t. Finally, in Section G.3, we describe how to simulate the reverse process across the window and then recover the original point, completing the proof of Theorem G.1.

G.1 Quantifying the critical window

We begin by comparing the density of an observation X_t started from one point $Y \in \mathcal{X}$ with the density obtained from a uniform initial point on the whole cube. Normalize the observation by its noise standard deviation:

$$Z = X_t/\sigma_t = \lambda_t Y + G, \qquad \lambda_t = e^{-t}/\sigma_t\,. \tag{57}$$

For a general parameter $u > 0$, write p_u for the density of $uU + G$ with U uniform on $\mathcal{X}$. Write ϕ_d for the standard Gaussian density, so

$$p_u(z) = \mathbb{E}_{Y \sim \mathrm{Unif}(\mathcal{X})} \phi_d(z - uY) = \phi_d(z) e^{-u^2 d/2} \prod_{i=1}^{d} \cosh(u z_i)\,. \tag{58}$$

Consequently, define the log-likelihood ratio of a point $y \in \mathcal{X}$ given an observation z as

$$\imath_t(y; z) = \log \frac{\phi_d(z - \lambda_t y)}{p_{\lambda_t}(z)} = \lambda_t \langle z, y\rangle - \sum_{i=1}^{d} \log\cosh(\lambda_t z_i)\,.$$

Intuitively, $\imath_t(y; z)$ measures how y stands out from the uniform distribution given the observation z. Under the joint distribution of $Y \sim \mathrm{Unif}(\mathcal{X})$ and $Z = \lambda_t Y + G$, this is a sum of d independent terms $\lambda_t z_i y_i - \log\cosh(\lambda_t z_i)$, each with mean $\mathsf{I}_{\mathsf{gauss}}(t)$. Below we will show that it concentrates around its mean, and that the mean is strictly decreasing with a controlled speed.

Lemma G.2. *Uniformly over $t \in I_0$ and $v \geq 0$,*

$$\mathop{\mathbb{P}}_{Y \sim \mathrm{Unif}(\mathcal{X}),\, Z = \lambda_t Y + G} [|\imath_t(Y; Z) - d\mathsf{I}_{\mathsf{gauss}}(t)| \geq v] \leqslant 2\exp\left(-cv^2/d\right)\,. \tag{59}$$

Moreover, there are constants $0 < c_I < C_I$ such that

$$c_I \leqslant -\mathsf{I}'_{\mathsf{gauss}}(t) \leqslant C_I, \qquad t \in I_0\,. \tag{60}$$

Proof. For the concentration bound, write

$$\imath_t(Y; Z) = \sum_{i=1}^{d} \xi_{t,i}\,, \qquad \xi_{t,i} = \lambda_t Z_i Y_i - \log\cosh(\lambda_t Z_i)\,.$$

Let $H_i = Y_i G_i$. Then $H_1, \ldots, H_d$ are independent standard Gaussians, and since $Z_i Y_i = \lambda_t + H_i$ and cosh is even,

$$\xi_{t,i} = f_t(H_i)\,, \qquad f_t(h) = \lambda_t(\lambda_t + h) - \log\cosh\big(\lambda_t(\lambda_t + h)\big)\,.$$

Moreover,

$$f_t'(h) = \lambda_t \left[1 - \tanh\big(\lambda_t(\lambda_t + h)\big)\right]\,.$$

Since λ_t is bounded for $t \in I_0$, $|f_t'(h)| \leqslant C$ uniformly in t and h. Hence the function $(h_1, \dots, h_d) \longmapsto \sum_{i=1}^d f_t(h_i)$ is $C\sqrt{d}$-Lipschitz. Gaussian concentration gives Eq. (59).

To check the derivative, consider one coordinate. Write $V \sim \mathrm{Unif}(\{-1, 1\}), W \sim N(0, 1)$ independent, and $u = \lambda_t$. Then

$$\mathsf{I}_{\mathsf{gauss}}(t) = I(V; uV + W) = u^2 - \mathbb{E}_W \log\cosh(u^2 + uW)\,.$$

Differentiate in u and use Gaussian integration by parts,

$$\frac{\mathrm{d}}{\mathrm{d}u} I(V; uV + W) = 2u - 2u\,\mathbb{E}\tanh H - u\,\mathbb{E}\mathrm{sech}^2 H, \qquad H = u^2 + uW\,.$$

Since $\mathbb{E}[V \mid uV + W] = \tanh(u(uV + W))$,

$$\mathbb{E}\tanh H = \mathbb{E}[V\,\mathbb{E}[V \mid uV + W]] = \mathbb{E}[\mathbb{E}[V \mid uV + W]^2] = \mathbb{E}\tanh^2 H\,.$$

Substituting this identity into the derivative, we get

$$\frac{\mathrm{d}}{\mathrm{d}u} I(V; uV + W) = u\,\mathbb{E}_W \mathrm{sech}^2(u^2 + uW)\,.$$

This derivative is continuous and strictly positive for $u > 0$. Both λ_t and $|\lambda_t'|$ are bounded above and away from zero on I_0, and $\lambda_t' < 0$. The chain rule now gives Eq. (60). □

For a fixed $\mathcal{C}$, denote the smallest probability of failing to recover $Y \sim q^{\mathcal{C}}$ from $\lambda_t Y + G$ by $\mathrm{err}_{\mathcal{C}}^{\mathsf{G}}(t)$:

$$\mathrm{err}_{\mathcal{C}}^{\mathsf{G}}(t) \coloneqq \inf_{\widehat{Y}} \mathop{\mathbb{P}}_{Y \sim q^{\mathcal{C}}}[\widehat{Y}(\lambda_t Y + G) \neq Y \mid \mathcal{C}]\,.$$

The following lemma quantifies our intuition for the log-likelihood ratio $\imath_t(Y; Z)$.

Lemma G.3. *For every $t > 0$ and $\gamma > 0$,*

$$\mathbb{E}_{\mathcal{C}}\mathsf{TV}(q_t^{\mathcal{C}}, \nu_t) \leqslant \mathbb{P}[\imath_t(Y; Z) > \log M - \gamma] + \tfrac{1}{2}e^{-\gamma/2}, \tag{61}$$

$$\mathbb{E}_{\mathcal{C}}\mathrm{err}_{\mathcal{C}}^{\mathsf{G}}(t) \leqslant \mathbb{P}[\imath_t(Y; Z) \leqslant \log M + \gamma] + e^{-\gamma}\,, \tag{62}$$

where both probabilities are taken with respect to $Y \sim \mathrm{Unif}(\mathcal{X})$ and $Z = \lambda_t Y + G$.

Proof. First consider bounding $\mathbb{E}_{\mathcal{C}}\mathsf{TV}(q_t^{\mathcal{C}}, \nu_t)$. Fix t, and write $L_y(z) = \phi_d(z - \lambda_t y)/p_{\lambda_t}(z)$. For every z, the average of $L_y(z)$ over the whole cube is 1. We have

$$\mathsf{TV}(q_t^{\mathcal{C}}, \nu_t) = \frac{1}{2}\int p_{\lambda_t}(z)\left|\frac{1}{M}\sum_{y \in \mathcal{C}} L_y(z) - 1\right|\mathrm{d}z\,,$$

and $M^{-1}\sum_{y\in\mathcal{C}} L_y(z)$ is the empirical average of $L_y(z)$. To control its fluctuation, truncate the likelihood ratio at $R = Me^{-\gamma}$ and define

$$\widetilde{L}_y(z) = L_y(z)\,\mathbb{1}[L_y(z) \leqslant R], \qquad \widetilde{L}_y^c(z) = L_y(z)\,\mathbb{1}[L_y(z) > R], \qquad \mu(z) = \mathbb{E}_{Y\sim\mathrm{Unif}(\mathcal{X})}\widetilde{L}_Y(z)\,.$$

We have the following decomposition:

$$2\mathbb{E}_{\mathcal{C}}\mathsf{TV}(q_t^{\mathcal{C}}, \nu_t) \leqslant \underbrace{\mathbb{E}_{\mathcal{C}}\int p_{\lambda_t}(z)\left|\frac{1}{M}\sum_{y\in\mathcal{C}}\widetilde{L}_y(z) - \mu(z)\right|\mathrm{d}z}_{①} + \underbrace{\mathbb{E}_{\mathcal{C}}\int p_{\lambda_t}(z)\Big(\frac{1}{M}\sum_{y\in\mathcal{C}}\widetilde{L}_y^c(z)\Big)\mathrm{d}z}_{②}$$
$$+ \underbrace{\mathbb{E}_{Y\sim\mathrm{Unif}(\mathcal{X})}\int p_{\lambda_t}(z)L_Y(z)\,\mathbb{1}[L_Y(z) > R]\mathrm{d}z}_{③}\,.$$

Fix z now. For ①, by Cauchy–Schwarz and the variance bound for sampling without replacement,

$$\mathbb{E}_{\mathcal{C}}\left|\frac{1}{M}\sum_{y\in\mathcal{C}}\widetilde{L}_y(z) - \mu(z)\right| \leqslant \sqrt{\frac{1}{M}\mathbb{E}_{Y\sim\mathrm{Unif}(\mathcal{X})}\widetilde{L}_Y(z)^2} \leqslant \sqrt{\frac{R}{M}} = e^{-\gamma/2}\,.$$

Here we used $\widetilde{L}_Y^2 \leqslant R\widetilde{L}_Y$ and $\mu(z) \leqslant 1$. Integrating over p_{λ_t} gives the same bound. Finally,

$$② = ③ = \mathbb{E}_{Y\sim\mathrm{Unif}(\mathcal{X})} \int p_{\lambda_t}(z) L_Y(z)\, \mathbb{1}[L_Y(z) > R]\, \mathrm{d}z = \mathbb{P}[\imath_t(Y;Z) > \log M - \gamma]\,,$$

and Eq. (61) follows.

Next we bound the Bayes recovery error. Given $\mathcal{C}$ and z, return the unique point $y \in \mathcal{C}$ with $L_y(z) > Me^{\gamma}$, if there is one. This rule succeeds if the initial point passes the test and all other points fail it. The probability that the initial point fails is the first term in Eq. (62). We will bound the probability that any other point passes the test.

Fix the initial point y and the observation z, and consider the randomness of $\mathcal{C}$. The remaining $M-1$ points are a uniform subset of $\mathcal{X} \setminus \{y\}$. Since $\mathbb{E}_{Y'\sim\mathrm{Unif}(\mathcal{X})} L_{Y'}(z) = 1$, by Markov's inequality we have

$$\mathop{\mathbb{P}}_{Y'\sim\mathrm{Unif}(\mathcal{X})}[L_{Y'}(z) > Me^{\gamma}] \leqslant M^{-1}e^{-\gamma}\,.$$

Removing y from the sampling population increases this bound by at most a factor of $2^d/(2^d - 1) < M/(M-1)$. A union bound over the remaining points gives the bound $e^{-\gamma}$, and Eq. (62) follows. □

Setting the parameters appropriately in the two estimates above characterizes the Gaussian critical window: above it the noised empirical measure is close to the reference distribution, while below it the original point can be recovered with high probability.

Proposition G.4 (Critical window)**.** *For every $K > 0$ there is $A_K > 0$ with the following property. For every $A \geq A_K$, there are $c_{K,A}, d_{K,A} > 0$, depending only on A, K, and $[\kappa_-, \kappa_+]$, such that the following holds whenever $d \geq d_{K,A}$ and $\log d \leqslant \Lambda \leqslant c_{K,A} d$. Set*

$$w = A\sqrt{\frac{\Lambda}{d}}\,. \tag{63}$$

Then

$$\begin{aligned} \mathbb{E}_{\mathcal{C}}\mathsf{TV}(q_t^{\mathcal{C}}, \nu_t) &\leqslant e^{-K\Lambda}\,, & t &\geq t^\star_{\mathsf{gauss}} + w\,,\\ \mathbb{E}_{\mathcal{C}}\mathrm{err}_{\mathcal{C}}^{\mathsf{G}}(t) &\leqslant e^{-K\Lambda}\,, & 0 &\leqslant t \leqslant t^\star_{\mathsf{gauss}} - w\,. \end{aligned}$$

Proof. By decreasing $c_{K,A}$ if necessary, we may assume $w \leqslant \min\{\tau_-/2, 1\}$, so that $t^\star_{\mathsf{gauss}} \pm w \in I_0$. Fix $\mathcal{C}$. Applying an additional forward step to both distributions cannot increase $\mathsf{TV}(q_t^{\mathcal{C}}, \nu_t)$. Also, an earlier observation can be used to simulate a later one, so $\mathrm{err}_{\mathcal{C}}^{\mathsf{G}}(t)$ cannot decrease with t. It is therefore enough to prove the claims at $t^\star_{\mathsf{gauss}} + w$ and $t^\star_{\mathsf{gauss}} - w$.

Above the critical window. Set $t = t^\star_{\mathsf{gauss}} + w$. Since $d\,\mathsf{I}_{\mathsf{gauss}}(t^\star_{\mathsf{gauss}}) = \log M$, Eq. (60) gives $d\,\mathsf{I}_{\mathsf{gauss}}(t) \leqslant \log M - c_I dw$. Take $\gamma = c_I dw/2$ in Eq. (61). The event $\imath_t(Y;Z) > \log M - \gamma$ requires a deviation for $\imath_t(Y;Z)$ of at least $c_I dw/2$ above the mean $d\,\mathsf{I}_{\mathsf{gauss}}(t)$. Lemma G.2 bounds its probability by $2e^{-cdw^2}$. The remaining term is at most $e^{-cdw} \lesssim e^{-cdw^2}$.

Below the critical window. Set $t = t^\star_{\mathsf{gauss}} - w$. Now $d\,\mathsf{I}_{\mathsf{gauss}}(t) \geq \log M + c_I dw$. With the same choice $\gamma = c_I dw/2$ in Eq. (62), the event $\imath_t(Y;Z) \leqslant \log M + \gamma$ requires a deviation of at least $c_I dw/2$ below the mean $d\,\mathsf{I}_{\mathsf{gauss}}(t)$. Its probability is again at most $2e^{-cdw^2}$, and the other term is at most $e^{-cdw} \lesssim e^{-cdw^2}$.

Finally, since $dw^2 = A^2\Lambda$, increasing A gives the desired exponent. The monotonicity noted at the beginning extends the bounds to the stated time ranges. □

For the rest of the section, let $w = A\sqrt{\Lambda/d}$ as in Eq. (63), where $A > 0$ is a constant to be fixed below. We may further decrease c_0 and increase d_0 later when necessary.

G.2 IDENTIFYING THE CRITICAL WINDOW

Having characterized the critical window, we now show how to identify its location using approximate Gaussian score oracle queries at samples from the reference distribution.

Define the density ratio of $q_t^{\mathcal{C}}$ to ν_t as

$$L_t(x) \coloneqq \frac{q_t^{\mathcal{C}}(x)}{\nu_t(x)}\,. \tag{64}$$

Given $x \in \mathbb{R}^d$, define the posterior means under $q^{\mathcal{C}}$ and under the uniform distribution on the whole cube by

$$m_t(x) \coloneqq \mathbb{E}[X_0 \mid X_t = x], \qquad m_t^0(x)_i \coloneqq \tanh\left(\frac{e^{-t}x_i}{\sigma_t^2}\right), \quad i \in [d]\,. \tag{65}$$

For $z = x/\sigma_t$, consider the difference between the two expected overlaps $\langle z, m_t(x)\rangle$ and $\langle z, m_t^0(x)\rangle$:

$$T_t(x) = \langle z, m_t(x) - m_t^0(x)\rangle\,. \tag{66}$$

Above the window, both distributions $q_t^{\mathcal{C}}$ and ν_t assign nearly the same density to most observations drawn from ν_t, and this gap is small. Below the window, a sample from ν_t is unlikely to be close to any point of $\mathcal{C}$ since $\mathcal{C}$ is exponentially sparse, so the gap is large. These two behaviors can be used to identify the critical window. We prove these two assertions before considering estimated scores.

As in the uniform diffusion proof, it is useful to express both the density ratio and the overlap difference through two log averages. For fixed $\mathcal{C}, z$, define

$$A_z(r) = \log \mathbb{E}_{Y\sim q^{\mathcal{C}}} e^{r\langle z,Y\rangle}, \qquad B_z(r) = \log \mathbb{E}_{Y\sim \mathrm{Unif}(\mathcal{X})} e^{r\langle z,Y\rangle} = \sum_{i=1}^{d} \log\cosh(rz_i)\,. \tag{67}$$

At $u = \lambda_t$ and $z = x/\sigma_t$, Bayes' rule gives

$$\log L_t(x) = A_z(u) - B_z(u), \qquad T_t(x) = A_z'(u) - B_z'(u)\,. \tag{68}$$

Thus $A_z - B_z$ compares the two densities, while its derivative is exactly the quantity used by our test.

Let $\pi_r(\cdot \mid z)$ be the conditional distribution of $U \sim \mathrm{Unif}(\mathcal{X})$ given $rU + G = z$. Then

$$\pi_r(y \mid z) = 2^{-d} e^{r\langle z,y\rangle - B_z(r)} = \prod_{i=1}^{d} \frac{e^{rz_iy_i}}{2\cosh(rz_i)}\,. \tag{69}$$

Define the *posterior information*

$$\mathcal{I}_z(r) \coloneqq \mathsf{KL}(\pi_r(\cdot \mid z) \,\|\, \mathrm{Unif}(\mathcal{X})) = d\log 2 - H(\pi_r(\cdot \mid z)) = rB_z'(r) - B_z(r)\,. \tag{70}$$

Indeed, Eq. (69) gives $\log(\pi_r(y \mid z)/2^{-d}) = r\langle z, y\rangle - B_z(r)$, and taking the posterior expectation gives the last identity in Eq. (70). Thus $\mathcal{I}_z(r)$ is the information about a uniformly random cube point contained in the observation z. The exponential of the posterior entropy is exactly $2^d e^{-\mathcal{I}_z(r)}$. When $r = \lambda_t$ and $Z \sim p_{\lambda_t}$, taking expectation of $\mathcal{I}_Z(\lambda_t)$ over Z gives the mutual information $d\,\mathsf{I}_{\mathsf{gauss}}(t)$. At the critical time $t^\star_{\mathsf{gauss}}$ this equals $\log M$, so the entropy scale equals $2^d/M$, the reciprocal of the fraction of the cube contained in $\mathcal{C}$.

The derivatives of these quantities also have direct meanings. Under $\pi_r(\cdot \mid z)$, the mean and variance of the overlap $\langle z, U\rangle$ are $B_z'(r)$ and $B_z''(r)$, and hence

$$\mathcal{I}_z'(r) = rB_z''(r) = r \operatorname*{\mathbb{V}}_{U\sim\pi_r(\cdot|z)}[\langle z, U\rangle]\,.$$

In particular, the posterior information increases with the signal strength r.

We need two elementary facts about the reference law. The first ensures enough overlap between p_u and p_v if u and v are close.

Lemma G.5. *For $u, v > 0$ and every measurable set $A \subseteq \mathbb{R}^d$,*

$$p_u(A) \leqslant e^{d(u-v)^2/2}\sqrt{p_v(A)}\,. \tag{71}$$

More generally, let $\mathcal{C}$ be a random subset independent of (U, G), with the same law under parameters u and v. For every measurable event $\mathcal{E}$ determined by $(\mathcal{C}, Z)$,

$$\mathop{\mathbb{P}}_{\mathcal{C},Z\sim p_u}[\mathcal{E}] \leqslant e^{d(u-v)^2/2}\sqrt{\mathop{\mathbb{P}}_{\mathcal{C},Z\sim p_v}[\mathcal{E}]}\,. \tag{72}$$

We defer the proof to Appendix J.

The second fact says that, for a typical sample from p_{λ_t}, the posterior information is close to its mean and changes at a speed of order d.

Lemma G.6. *For every $K > 0$, there are constants $C_K, c_K, d_K > 0$, depending only on $[\kappa_-, \kappa_+]$ and K, such that the following holds whenever $d \geq d_K$, $\log d \leqslant \Lambda \leqslant c_K d$, and $t \in I_0$. With probability at least $1 - e^{-K\Lambda}$ for $Z \sim p_{\lambda_t}$,*

$$\|Z\|_2^2 \leqslant Cd, \tag{73}$$

$$cd \leqslant B_Z''(r) \leqslant Cd, \tag{74}$$

$$|\mathcal{I}_Z(\lambda_t) - d\mathsf{I}_{\mathsf{gauss}}(t)| \leqslant C_K\sqrt{d\Lambda}\,. \tag{75}$$

The bounds on B_Z'' hold uniformly for $r \in [\lambda_{\tau_+ +1}/2, 2\lambda_{\tau_-/2}]$, and c, C depend only on $[\kappa_-, \kappa_+]$.

We defer the proof to Appendix J.

We can now use the above ingredients to prove that the statistic T_t can be used to distinguish whether one is above or below the critical window:

Proposition G.7. *There are constants $a_0, b_0 > 0$, depending only on $[\kappa_-, \kappa_+]$, with $b_0 > 4a_0$, such that for every $K > 0$ there is $A_K > 0$ with the following property. For every $A \geq A_K$, there are $c_{K,A}, d_{K,A} > 0$, depending only on A, K, and $[\kappa_-, \kappa_+]$, such that the following holds whenever $d \geq d_{K,A}$ and $\log d \leqslant \Lambda \leqslant c_{K,A} d$:*

$$\mathop{\mathbb{P}}_{\mathcal{C},X\sim\nu_t}[T_t(X) < -a_0 dw] \leqslant e^{-K\Lambda}, \qquad t \geq t^\star_{\mathsf{gauss}} + 2w, \tag{76}$$

$$\mathop{\mathbb{P}}_{\mathcal{C},X\sim\nu_t}[T_t(X) > -b_0 dw] \leqslant e^{-K\Lambda}, \qquad t \leqslant t^\star_{\mathsf{gauss}} - 2w\,, \tag{77}$$

uniformly over $t \in I_0$.

Proof. Write $u = \lambda_t$ and $z = X/\sigma_t$.

High-noise regime ($t \geq t^\star_{\mathsf{gauss}} + 2w$). Choose a small constant $\theta > 0$, set $t^+ = t + \theta w$, and write $v = \lambda_{t^+}$. Since $-\lambda_t'$ is bounded above and below on the relevant compact interval, $u - v \asymp \theta w$. In particular, t and t^+ are both above $t^\star_{\mathsf{gauss}} + w$. By Eq. (68), define the two density ratios, as functions of the common normalized observation z, by

$$L_u(z) = e^{A_z(u) - B_z(u)}, \qquad L_v(z) = e^{A_z(v) - B_z(v)}\,.$$

Proposition G.4 and Markov's inequality give

$$\mathop{\mathbb{P}}_{\mathcal{C},Z\sim p_u}[|L_u(Z) - 1| > 1/2] + \mathop{\mathbb{P}}_{\mathcal{C},Z\sim p_v}[|L_v(Z) - 1| > 1/2] \leqslant Ce^{-cdw^2}\,.$$

Lemma G.5 transfers the bound for L_v from p_v to p_u. Since $d(u-v)^2 \leqslant C\theta^2 dw^2$, we may choose θ small enough that, under $\mathcal{C} \sim \mathrm{Unif}\big(\binom{\mathcal{X}}{M}\big)$ and $Z \sim p_u$, both $L_u(Z)$ and $L_v(Z)$ belong to $[1/2, 3/2]$ except with probability Ce^{-cdw^2}. Lemma G.6 gives $B_Z''(r) \leqslant Cd$ uniformly for $r \in [v, u]$ outside another event of probability $e^{-K\Lambda}$, for sufficiently large d.

Fix $\mathcal{C}, z$ for which these three conclusions hold. Since A_z is convex,

$$\begin{aligned} T_t(x) &= A_z'(u) - B_z'(u) \\ &\geq \frac{A_z(u) - A_z(v)}{u - v} - B_z'(u) \\ &= \frac{\log L_u(z) - \log L_v(z)}{u - v} + \frac{B_z(u) - B_z(v)}{u - v} - B_z'(u) \\ &\geq -\frac{C \log 3}{\theta w} - Cd\theta w \,. \end{aligned}$$

The last step uses $u - v \asymp \theta w$, Eq. (74), and Taylor's formula. For any prescribed $a_0 > 0$, further choose θ small enough so that the last term is at least $-a_0 dw/2$, and then choose A large enough so that the first term is at least $-a_0 dw/2$, using $dw^2 = A^2\Lambda$. Increasing A also makes the total exceptional probability at most $e^{-K\Lambda}$.

Low-noise regime ($t \leqslant t^\star_{\mathsf{gauss}} - 2w$). Apply Lemma G.6 with exponent $K+1$ to $Z \sim p_u$, and fix an observation z satisfying Eq. (75). Since $\mathsf{I}_{\mathsf{gauss}}(t) - \mathsf{I}_{\mathsf{gauss}}(t^\star_{\mathsf{gauss}}) \geq 2c_I w$ and $d\mathsf{I}_{\mathsf{gauss}}(t^\star_{\mathsf{gauss}}) = \log M$, choosing A sufficiently large gives

$$\mathcal{I}_z(u) \geq \log M + c_I dw \,. \tag{78}$$

We show directly that $\mathcal{C}$ is unlikely to contain even one point y whose overlap $\langle z, y\rangle$ reaches the reference posterior mean $B_z'(u)$ up to a small gap $b_0 dw$. Choose $b_0 > 0$ so that $ub_0 \leqslant c_I/2$ uniformly in u. By Markov's inequality and Eq. (70),

$$\begin{aligned} \mathop{\mathbb{P}}_{Y \sim \mathrm{Unif}(\mathcal{X})}[\langle z, Y\rangle \geq B_z'(u) - b_0 dw] &\leqslant e^{-u(B_z'(u) - b_0 dw)} \mathbb{E}_{Y \sim \mathrm{Unif}(\mathcal{X})} e^{u\langle z, Y\rangle} \\ &= \exp\{-\mathcal{I}_z(u) + ub_0 dw\} \\ &\leqslant M^{-1} e^{-c_I dw/2} \,. \end{aligned}$$

By a union bound over the M points in $\mathcal{C}$, except with probability $e^{-c_I dw/2}$, every $y \in \mathcal{C}$ has overlap less than $B_z'(u) - b_0 dw$. Their posterior-weighted average is $A_z'(u)$, so on this event

$$T_t(x) = A_z'(u) - B_z'(u) \leqslant -b_0 dw \,.$$

Adding the probability $e^{-(K+1)\Lambda}$ that Eq. (78) fails, and choosing A large enough that $e^{-c_I dw/2} \leqslant e^{-(K+1)\Lambda}$, which is possible since $dw \geq dw^2 = A^2\Lambda$, we obtain Eq. (77). Finally, choose a_0 in the first part small enough that $b_0 > 4a_0$. □

We now replace the posterior mean by the estimate obtained from the score. Define its clipped version coordinatewise by

$$\widetilde{m}_t(x)_i \coloneqq \min\left\{1, \max\left\{-1, e^t x_i + e^t \sigma_t^2 \widehat{s}_t(x)_i\right\}\right\}, \qquad i \in [d] \,. \tag{79}$$

Using this vector, define the clipped approximate score and the statistic available to the algorithm by

$$\widetilde{s}_t(x) \coloneqq \frac{e^{-t}\widetilde{m}_t(x) - x}{\sigma_t^2}, \qquad \widehat{T}_t(x) \coloneqq \langle x/\sigma_t, \widetilde{m}_t(x) - m_t^0(x)\rangle \,. \tag{80}$$

Both quantities can be computed from one approximate score oracle query at x.

The following lemma shows that the estimated overlap difference $\widehat{T}_t$ is indeed a good approximation of the true overlap difference T_t.

Lemma G.8. *For every fixed $\mathcal{C}$ and $t \in I_0$,*

$$\mathbb{E}_{q_t^{\mathcal{C}}} \|\widetilde{m}_t - m_t\|_2^2 \leqslant C\varepsilon_{\mathsf{gauss}}(t)^2, \qquad \mathbb{E}_{q_t^{\mathcal{C}}} \|\widetilde{s}_t - s_t\|_2^2 \leqslant C\varepsilon_{\mathsf{gauss}}(t)^2 \,. \tag{81}$$

Moreover, whenever $\|x/\sigma_t\|_2 \leqslant C_1\sqrt{d}$,

$$|\widehat{T}_t(x) - T_t(x)| \leqslant C_1\sqrt{d}\, \|\widetilde{m}_t(x) - m_t(x)\|_2 \,. \tag{82}$$

We defer the proof to Appendix J.

The score assumption controls errors on observations from $q_t^{\mathcal{C}}$, while our queries use ν_t. Above the window, total variation already compares the two distributions. Near the window, we need the following weaker assertion: a sample from ν_t is unlikely to have extremely small density under $q_t^{\mathcal{C}}$.

Lemma G.9. *For every $A, L, K > 0$, there are constants $C_{A,L,K}, c_{A,L,K}, d_{A,L,K} > 0$, depending only on A, L, K and $[\kappa_-, \kappa_+]$, such that whenever $d \geq d_{A,L,K}$ and $\log d \leqslant \Lambda \leqslant c_{A,L,K} d$,*

$$\mathop{\mathbb{P}}_{\mathcal{C}, X \sim \nu_t}[L_t(X) < e^{-C_{A,L,K}\Lambda}] \leqslant e^{-K\Lambda}\,, \tag{83}$$

uniformly over $|t - t^\star_{\mathsf{gauss}}| \leqslant Lw$.

We defer the proof to Appendix J.

The above lemmas allow us to use the estimated overlap difference $\widehat{T}_t$ to identify the critical window.

Proposition G.10. *There are constants $0 < a < b$, depending only on $[\kappa_-, \kappa_+]$, such that for every $K > 0$ there is $A_K > 0$ with the following property. For every $A \geq A_K$, there are $B_{K,A}, c_{K,A}, d_{K,A} > 0$ such that the following holds whenever $d \geq d_{K,A}$, $\log d \leqslant \Lambda \leqslant c_{K,A} d$, $B \geq B_{K,A}$, and the score oracle satisfies Eq. (55):*

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathcal{C}, X \sim \nu_t}[\widehat{T}_t(X) < -adw] \leqslant e^{-K\Lambda}, && t \geq t^\star_{\mathsf{gauss}} + 2w,\\
&\mathop{\mathbb{P}}_{\mathcal{C}, X \sim \nu_t}[\widehat{T}_t(X) > -bdw] \leqslant e^{-K\Lambda}, && t^\star_{\mathsf{gauss}} - 3w \leqslant t \leqslant t^\star_{\mathsf{gauss}} - 2w\,,
\end{aligned}$$

uniformly over $t \in I_0$.

Proof. Take a_0, b_0 from Proposition G.7, and put $a = 2a_0$ and $b = b_0 - a_0$. These satisfy $a < b$. We will show that replacing T_t by $\widehat{T}_t$ changes it by at most $a_0 dw$ with sufficiently high probability in each of the two ranges.

Fix $\mathcal{C}$ and t, and let $c > 0$ be a small constant to be fixed below. By Lemma G.8 and Markov's inequality,

$$\mathop{\mathbb{P}}_{X \sim q_t^{\mathcal{C}}}[\|\widetilde{m}_t(X) - m_t(X)\|_2 > c\sqrt{d}\, w] \leqslant \frac{Ce^{-2B\Lambda}}{dw^2}\,. \tag{84}$$

Choose c small enough that Eq. (82) bounds the overlap error by $a_0 dw$ whenever this event does not occur and $\|X/\sigma_t\|_2 \leqslant C\sqrt{d}$. The latter condition fails under ν_t with probability at most $e^{-(K+3)\Lambda}$, by Lemma G.6.

High-noise regime ($t \geq t^\star_{\mathsf{gauss}} + 2w$). Replacing $q_t^{\mathcal{C}}$ by ν_t in Eq. (84) adds at most $\mathsf{TV}(q_t^{\mathcal{C}}, \nu_t)$. Average over $\mathcal{C}$ and apply Proposition G.4 with exponent $K+3$. Together with the $\|X/\sigma_t\|_2 \leqslant C\sqrt{d}$ probability bound, this gives

$$\mathop{\mathbb{P}}_{\mathcal{C}, X \sim \nu_t}[|\widehat{T}_t(X) - T_t(X)| > a_0 dw] \leqslant \frac{Ce^{-2B\Lambda}}{dw^2} + 2e^{-(K+3)\Lambda}\,.$$

Increasing B makes this at most $e^{-(K+1)\Lambda}$. Outside this event and the event in Eq. (76), we have $\widehat{T}_t \geq -2a_0 dw = -adw$.

Low-noise regime ($t^\star_{\mathsf{gauss}} - 3w \leqslant t \leqslant t^\star_{\mathsf{gauss}} - 2w$). We apply Lemma G.9 with $L = 3$ and exponent $K + 3$, and denote its density exponent by C_1. On the event $L_t(X) = q_t^{\mathcal{C}}(X)/\nu_t(X) \geq e^{-C_1\Lambda}$, integration under ν_t is bounded by $e^{C_1\Lambda}$ times integration under $q_t^{\mathcal{C}}$. Then following the same steps as in the high-noise regime, we obtain

$$\mathop{\mathbb{P}}_{\mathcal{C}, X \sim \nu_t}[|\widehat{T}_t(X) - T_t(X)| > a_0 dw] \leqslant \frac{Ce^{-(2B - C_1)\Lambda}}{dw^2} + 2e^{-(K+3)\Lambda}\,.$$

Choose B after C_1 so that this is at most $e^{-(K+1)\Lambda}$. Outside this event and the event in Eq. (77), we have $\widehat{T}_t \leqslant -(b_0 - a_0)dw = -bdw$.

Finally, applying Proposition G.7 with exponent $K+2$ and adding the error probabilities proves both assertions. Here A_K is the larger of the thresholds required by Proposition G.4 at exponent $K+3$ and by Proposition G.7 at exponent $K+2$. For every fixed $A \geq A_K$, apply Lemma G.9 with $L=3$ and exponent $K+3$, choose $B_{K,A}$ sufficiently large in terms of its density exponent, and then choose $c_{K,A}$ sufficiently small and $d_{K,A}$ sufficiently large so that all three results apply. □

From now on, fix $A \geq A_{12}$ and choose $B \geq B_{12,A}$. We may decrease c_0 and increase d_0 so that $c_0 \leqslant c_{12,A}$, $d_0 \geq d_{12,A}$, and $w \leqslant \min \tau_-/20, 1/20$. We further decrease c_0 and increase d_0 below whenever necessary.

The two bounds tell us how to search. We test decreasing times using samples from the reference distribution at each time. The test should not stop above $t^\star_{\mathsf{gauss}} + 2w$, and it should stop by the time it reaches $[t^\star_{\mathsf{gauss}} - 3w, t^\star_{\mathsf{gauss}} - 2w]$. Its behavior elsewhere is irrelevant.

Algorithm 3: Finding the Gaussian diffusion critical window

Input: d, $[\kappa_-, \kappa_+]$, w, Λ, and query access to the score oracle $\widehat{s}$.
Output: An estimate $\widehat{t}$ of the critical time $t^\star_{\mathsf{gauss}}$.

Let $\mathcal{T} = \{t_1 > t_2 > \cdots > t_m\}$ be a decreasing grid covering $[\tau_- - 3w, \tau_+ + 3w]$ with mesh $w/20$;
Set $R \leftarrow 2\lceil C\Lambda \rceil + 1$;
for $t \in \mathcal{T}$ **do**
  for $j \leftarrow 1$ **to** R **do**
    Draw independent $U^{(j)} \sim \mathrm{Unif}(\mathcal{X})$ and $G^{(j)} \sim N(0, I_d)$;
    Set $X^{(j)} \leftarrow e^{-t}U^{(j)} + \sigma_t G^{(j)}$;
    Query $\widehat{s}_t(X^{(j)})$ and compute $\widehat{T}_t(X^{(j)})$ using Eqs. (79)–(80);
  if $\mathrm{median}\{\widehat{T}_t(X^{(1)}), \ldots, \widehat{T}_t(X^{(R)})\} < -(a+b)dw/2$ **then**
    return t;
return τ_-;

We can now conclude the main result of this subsection: scanning from high to low noise and stopping when the median statistic becomes sufficiently negative locates the critical window.

Proposition G.11. *Under the hypotheses of Theorem G.1, Algorithm 3 uses $O(\Lambda/w)$ score oracle queries. For at least a $1 - \delta/4$ fraction of $\mathcal{C}$,*

$$\mathbb{P}[t^\star_{\mathsf{gauss}} - 3w \leqslant \widehat{t} \leqslant t^\star_{\mathsf{gauss}} + 2w \mid \mathcal{C}] \geq 1 - \varepsilon/8\,. \tag{85}$$

We defer the proof to Appendix J.

G.3 Sampling after locating the window

Once the search has returned $\widehat{t}$, we will start reverse simulation at $t_+ = \widehat{t} + 4w$ and stop at $t_- = \widehat{t} - 4w$. At each step, we hold the estimated score fixed and solve the remaining linear SDE exactly. If a step goes from time r to time $r - h$, its update is

$$Y_{\mathrm{next}} = e^h Y + 2(e^h - 1)\widetilde{s}_r(Y) + \sqrt{e^{2h} - 1}\, G, \qquad G \sim N(0, I_d)\,. \tag{86}$$

We need to bound the error from holding the score fixed for time h. A short interval alone is not enough: near the critical time, the posterior can change rapidly. The useful fact is that the total increase of its variance is at most d. We will charge the error on each step to the increase during that step and then sum these increases.

For a fixed $\mathcal{C}$, write

$$\Sigma_t(x) = \mathsf{Cov}(X_0 \mid X_t = x), \qquad V(t) = \mathbb{E}_{X_t \sim q_t^{\mathcal{C}}} \mathrm{Tr}\, \Sigma_t(X_t)\,. \tag{87}$$

Thus $V(t)$ is the mean squared error of the posterior mean. The following identity makes the preceding plan precise.

Lemma G.12. *Fix $\mathcal{C}$. For $t > 0$,*

$$V'(t) = 2e^{-2t}\sigma_t^{-4}\mathbb{E}_{q_t^{\mathcal{C}}} \operatorname{Tr}(\Sigma_t^2) \geq 0, \qquad 0 \leqslant V(t) \leqslant d\,. \tag{88}$$

We defer the proof to Appendix J.

We next bound the change of the score along one exact reverse step. Stating this separately keeps the time-discretization calculation apart from the score-estimation calculation.

Lemma G.13. *Fix $\mathcal{C}$ and $r - h, r \in I_0$, where $0 < h \leqslant 1$. Let $(Y_u)_{0\leqslant u\leqslant h}$ follow the exact reverse process from $q_r^{\mathcal{C}}$ to $q_{r-h}^{\mathcal{C}}$. Then*

$$\int_0^h \mathbb{E}\|s_{r-u}(Y_u) - s_r(Y_0)\|_2^2 \,\mathrm{d}u \leqslant C\left(dh^2 + h(V(r) - V(r-h))\right)\,. \tag{89}$$

We defer the proof to Appendix J.

We now combine the preceding score-change estimate with Girsanov's theorem on each grid interval, and then sum the errors over the grid.

Proposition G.14. *Fix $\mathcal{C}$ and a grid $t_- = r_0 < r_1 < \cdots < r_N = t_+$ in I_0. Put $h_j = r_j - r_{j-1}$, and suppose $\max_j h_j \leqslant h_0$, for a sufficiently small constant $h_0 > 0$ depending only on I_0. Initialize Eq. (86) from $q_{t_+}^{\mathcal{C}}$, and let $\widehat{q}_{t_-}^{\mathcal{C}}$ be its endpoint law. Then*

$$\mathsf{KL}(q_{t_-}^{\mathcal{C}} \parallel \widehat{q}_{t_-}^{\mathcal{C}}) \leqslant C\sum_{j=1}^{N}\left(dh_j^2 + h_j(V(r_j) - V(r_{j-1})) + h_j\mathbb{E}_{q_{r_j}^{\mathcal{C}}}\|\widetilde{s}_{r_j} - s_{r_j}\|_2^2\right). \tag{90}$$

In particular, on an equal grid with $\Delta = t_+ - t_-$,

$$\mathsf{KL}(q_{t_-}^{\mathcal{C}} \parallel \widehat{q}_{t_-}^{\mathcal{C}}) \leqslant C\left(\frac{d\Delta^2}{N} + \frac{\Delta}{N}(V(t_+) - V(t_-)) + \sum_{j=1}^{N}\frac{\Delta}{N}\mathbb{E}_{q_{r_j}^{\mathcal{C}}}\|\widetilde{s}_{r_j} - s_{r_j}\|_2^2\right)\,. \tag{91}$$

Proof. Fix $\mathcal{C}$ throughout. We compare the conditional path laws on each interval before summing over the grid.

One grid interval. Fix a grid interval $[r-h, r]$ and condition on $Y_0 = y$. Let $\mathsf{P}_{r,h}^{y}$ denote the path law of the exact reverse process

$$\mathrm{d}Y_u = \big(Y_u + 2s_{r-u}(Y_u)\big)\,\mathrm{d}u + \sqrt{2}\,\mathrm{d}B_u, \qquad Y_0 = y\,,$$

and let $\widetilde{\mathsf{P}}_{r,h}^{y}$ denote the path law of

$$\mathrm{d}\widetilde{Y}_u = \big(\widetilde{Y}_u + 2\widetilde{s}_r(y)\big)\,\mathrm{d}u + \sqrt{2}\,\mathrm{d}B_u, \qquad \widetilde{Y}_0 = y\,.$$

The time-h endpoint of the latter process is exactly the update in Eq. (86).

By Girsanov's theorem, we have

$$\begin{aligned}\mathsf{KL}(\mathsf{P}_{r,h}^{y} \parallel \widetilde{\mathsf{P}}_{r,h}^{y}) &= \frac{1}{4}\mathbb{E}_{\mathsf{P}_{r,h}^{y}}\int_0^h \|2s_{r-u}(Y_u) - 2\widetilde{s}_r(y)\|_2^2 \,\mathrm{d}u \\ &= \int_0^h \mathbb{E}\Big[\|s_{r-u}(Y_u) - \widetilde{s}_r(y)\|_2^2 \,\Big|\, Y_0 = y\Big]\,\mathrm{d}u\,.\end{aligned}$$

Averaging over $Y_0 \sim q_r^{\mathcal{C}}$ therefore gives

$$\mathbb{E}_{Y_0\sim q_r^{\mathcal{C}}}\mathsf{KL}(\mathsf{P}_{r,h}^{Y_0} \parallel \widetilde{\mathsf{P}}_{r,h}^{Y_0}) = \int_0^h \mathbb{E}\,\|s_{r-u}(Y_u) - \widetilde{s}_r(Y_0)\|_2^2 \,\mathrm{d}u\,. \tag{92}$$

The change-of-measure condition. We briefly verify Novikov's condition

$$\mathbb{E}\exp\left(\int_0^h \|s_{r-u}(Y_u)-\widetilde{s}_r(Y_0)\|_2^2\,\mathrm{d}u\right)<\infty$$

used above in Girsanov's theorem. By Tweedie's formula and the clipping in Eqs. (79)–(80),

$$\|s_t(x)\|_2+\|\widetilde{s}_t(x)\|_2\leqslant C(\|x\|_2+\sqrt{d}),\qquad t\in I_0\,.$$

Hence, along the exact reverse process,

$$\|s_{r-u}(Y_u)-\widetilde{s}_r(Y_0)\|_2^2\leqslant C\big(\|Y_u\|_2^2+\|Y_0\|_2^2+d\big)\,. \tag{93}$$

Under this process $Y_u\sim q^{\mathcal{C}}_{r-u}$, and every $q^{\mathcal{C}}_t$, $t\in I_0$, is a mixture of Gaussians $N(e^{-t}y,\sigma_t^2 I_d)$, $y\in\mathcal{C}$. Since their variances are uniformly bounded and $\|e^{-t}y\|_2\leqslant\sqrt{d}$, there is a constant $c>0$, depending only on I_0, such that

$$\sup_{\mathcal{C},t\in I_0}\mathbb{E}_{X\sim q^{\mathcal{C}}_t}e^{c\|X\|_2^2}<\infty\,. \tag{94}$$

Indeed, this follows from the quadratic exponential moment of a Gaussian, which is finite for $c<\frac{1}{2\sup_{t\in I_0}\sigma_t^2}$ after decreasing c if necessary.

Moreover, Jensen's inequality gives

$$\exp\left(C\int_0^h\|Y_u\|_2^2\,\mathrm{d}u\right)\leqslant\frac{1}{h}\int_0^h e^{Ch\|Y_u\|_2^2}\,\mathrm{d}u\,.$$

Combining this with Eq. (94) and Cauchy–Schwarz, we get

$$\mathbb{E}\exp\left(C\int_0^h\big(\|Y_u\|_2^2+\|Y_0\|_2^2+d\big)\,\mathrm{d}u\right)<\infty$$

whenever $h\leqslant h_0$ for a sufficiently small constant $h_0>0$ depending only on I_0. By Eq. (93), Novikov's condition follows.

Bounding the one-step error. We now return to Eq. (92). Adding and subtracting $s_r(Y_0)$, we separate the error caused by freezing the exact score from the error in the score estimate:

$$\int_0^h\mathbb{E}\,\|s_{r-u}(Y_u)-\widetilde{s}_r(Y_0)\|_2^2\,\mathrm{d}u\leqslant 2\int_0^h\mathbb{E}\,\|s_{r-u}(Y_u)-s_r(Y_0)\|_2^2\,\mathrm{d}u+2h\,\mathbb{E}_{q_r^{\mathcal{C}}}\|s_r-\widetilde{s}_r\|_2^2\,.$$

By Lemma G.13,

$$\int_0^h\mathbb{E}\,\|s_{r-u}(Y_u)-s_r(Y_0)\|_2^2\,\mathrm{d}u\leqslant C\big(dh^2+h(V(r)-V(r-h))\big)\,.$$

Therefore we get the following bound on the KL divergence between the exact and approximate path laws on a short interval:

$$\mathbb{E}_{Y_0\sim q_r^{\mathcal{C}}}\mathsf{KL}(\mathsf{P}^{Y_0}_{r,h}\parallel\widetilde{\mathsf{P}}^{Y_0}_{r,h})\leqslant C\big(dh^2+h(V(r)-V(r-h))\big)+2h\,\mathbb{E}_{q_r^{\mathcal{C}}}\|s_r-\widetilde{s}_r\|_2^2\,.$$

Summing over the whole grid. Let $\mathsf{P}^{\mathrm{path}}$ and $\widetilde{\mathsf{P}}^{\mathrm{path}}$ denote the path laws obtained by concatenating, respectively, the exact and approximate reverse processes over the grid intervals. The chain rule implies

$$\mathsf{KL}(\mathsf{P}^{\mathrm{path}}\parallel\widetilde{\mathsf{P}}^{\mathrm{path}})=\sum_{j=1}^N\mathbb{E}_{\mathsf{P}^{\mathrm{path}}}\mathsf{KL}(\mathsf{P}^{Y_{r_j}}_{r_j,h_j}\parallel\widetilde{\mathsf{P}}^{Y_{r_j}}_{r_j,h_j})\,.$$

Under the exact reverse law, $Y_{r_j}\sim q^{\mathcal{C}}_{r_j}$. Applying the preceding one-step bound on each interval therefore yields

$$\mathsf{KL}(\mathsf{P}^{\mathrm{path}}\parallel\widetilde{\mathsf{P}}^{\mathrm{path}})\leqslant C\sum_{j=1}^N\big[dh_j^2+h_j\big(V(r_j)-V(r_{j-1})\big)\big]+2\sum_{j=1}^N h_j\,\mathbb{E}_{q^{\mathcal{C}}_{r_j}}\big\|s_{r_j}-\widetilde{s}_{r_j}\big\|_2^2\,.$$

The endpoint is a measurable function of the full path, so the data processing inequality gives the same upper bound for $\mathsf{KL}(q^{\mathcal{C}}_{t_-} \,\|\, \widehat{q}^{\mathcal{C}}_{t_-})$. This proves Eq. (90).

For an equal grid, $h_j = \Delta/N$ with $\Delta = t_+ - t_-$. Hence

$$\sum_{j=1}^{N} dh_j^2 = \frac{d\Delta^2}{N}, \qquad \sum_{j=1}^{N} h_j\big(V(r_j) - V(r_{j-1})\big) = \frac{\Delta}{N}\big(V(t_+) - V(t_-)\big)\,.$$

Substituting these identities into the preceding bound gives Eq. (91). □

At the lower endpoint, we return the coordinatewise signs of $\widetilde{m}_t(x)$. This works for a reason stronger than a coordinatewise error estimate: most approximate samples below the window place almost all posterior mass on a single point of $\mathcal{C}$.

Lemma G.15. *Fix $\mathcal{C}$ and define $D_t(x) := \mathrm{sgn}(\widetilde{m}_t(x))$. If $X_0 \sim q^{\mathcal{C}}$ and X_t is obtained from X_0 through the Gaussian diffusion process, then*

$$\mathbb{P}[D_t(X_t) \neq X_0 \mid \mathcal{C}] \leqslant 5\mathrm{err}^{\mathsf{G}}_{\mathcal{C}}(t) + \mathop{\mathbb{P}}_{X\sim q^{\mathcal{C}}_t}[\|\widetilde{m}_t(X) - m_t(X)\|_2 > 1/4]\,.$$

Proof. Fix $X_t = x$, let $y_*(x) \in \mathcal{C}$ maximize the posterior probability $\mathbb{P}[X_0 = y \mid X_t = x]$, and call this probability $p_*(x)$. If $p_*(x) \geq 3/4$, every coordinate of $m_t(x)$ has the sign of $y_*(x)$ and magnitude at least $1/2$. An error at most $1/4$ in the Euclidean norm of the estimated posterior mean preserves all these signs. On these two events, $D_t(x) = y_*(x)$.

Averaging over X_t, we get $\mathbb{E}[1 - p_*(X_t) \mid \mathcal{C}] = \mathrm{err}^{\mathsf{G}}_{\mathcal{C}}(t)$. Markov's inequality bounds the probability of $p_*(X_t) < 3/4$ by $4\mathrm{err}^{\mathsf{G}}_{\mathcal{C}}(t)$. Add the probability of an error greater than $1/4$ in the posterior mean estimate to obtain the claim. □

Algorithm 4: Sampling from the random empirical measure with Gaussian diffusion

Input: d, $[\kappa_-, \kappa_+]$, target accuracy ε, failure probability δ, and a score oracle satisfying Eq. (55).
Output: A sample from the approximate output law $\widehat{q}^{\mathcal{C}}$.

Set Λ as in Eq. (54) and w as in Eq. (63);
Run Algorithm 3 to obtain $\widehat{t}$;
Set $t_+ \leftarrow \widehat{t} + 4w$, $t_- \leftarrow \widehat{t} - 4w$, and $\Delta \leftarrow 8w$;
Set $N \leftarrow \lceil C(d\Delta + d\Delta^2)/\varepsilon^2 \rceil$ and $r_j \leftarrow t_- + j\Delta/N$ for $0 \leqslant j \leqslant N$;
Draw $Y_N \sim \nu_{t_+}$ independently of Algorithm 3;
for $j = N, N-1, \ldots, 1$ **do**
    Query $\widehat{s}_{r_j}(Y_j)$ and compute $\widetilde{s}_{r_j}(Y_j)$ using Eqs. (79)–(80);
    Draw $G_j \sim N(0, I_d)$ independently and update Y_{j-1} by Eq. (86) with $h = \Delta/N$;
Query $\widehat{s}_{t_-}(Y_0)$ and compute $\widetilde{m}_{t_-}(Y_0)$ using Eq. (79);
return $\mathrm{sgn}(\widetilde{m}_{t_-}(Y_0))$;

Proof of Theorem G.1. We analyze Algorithm 4, allowing for errors in the search, the reverse steps, and the final rounding.

We first apply Proposition G.4 at the deterministic times $t^\star_{\mathsf{gauss}} + w$ and $t^\star_{\mathsf{gauss}} - w$. Markov's inequality, with exponent 12 in that proposition, removes at most a $\delta/2$ fraction of subsets and gives that for the remaining fraction of $\mathcal{C}$,

$$\begin{aligned} \mathsf{TV}(q^{\mathcal{C}}_t, \nu_t) &\leqslant \varepsilon/32, & t &\geq t^\star_{\mathsf{gauss}} + w, \\ \mathrm{err}^{\mathsf{G}}_{\mathcal{C}}(t) &\leqslant \varepsilon/160, & t &\leqslant t^\star_{\mathsf{gauss}} - w\,. \end{aligned} \tag{95}$$

Monotonicity extends the two bounds from the fixed two times to the two full ranges. The constant $\varepsilon/160$ is chosen so that $5\mathrm{err}^{\mathsf{G}}_{\mathcal{C}}(t) \leqslant \varepsilon/32$ in Lemma G.15 below. Proposition G.11 further removes

at most another $\delta/4$ fraction of $\mathcal{C}$ where the window search might fail. Fix a $\mathcal{C}$ satisfying all these properties for the remainder of the proof.

With conditional probability at least $1 - \varepsilon/8$, the critical window search returns $\widehat{t} \in [t^\star_{\mathsf{gauss}} - 3w, t^\star_{\mathsf{gauss}} + 2w]$. Now, condition on any such value $\widehat{t}$. Then $t_+ \geq t^\star_{\mathsf{gauss}} + w$, $t_- \leqslant t^\star_{\mathsf{gauss}} - w$, and $t_+ - t_- = \Delta = 8w$. The choices of c_0, d_0 ensure that the endpoints lie in I_0, even if the search fails. All samples used after the search are fresh.

First initialize the reverse steps from $q^{\mathcal{C}}_{t_+}$. Lemma G.12 gives $V(t_+) - V(t_-) \leqslant d$, and Lemma G.8 bounds the total score term by $C\Delta e^{-2B\Lambda}$. Proposition G.14 therefore gives

$$\mathsf{KL}(q^{\mathcal{C}}_{t_-} \parallel \widehat{q}^{\mathcal{C}}_{t_-}) \leqslant C\left(\frac{d\Delta^2 + d\Delta}{N} + \Delta e^{-2B\Lambda}\right) \leqslant \varepsilon^2/128\,.$$

The last inequality follows by choosing the constant in N and then B sufficiently large; the same choice ensures $\Delta/N \leqslant h_0$. Pinsker's inequality bounds $\mathsf{TV}(q^{\mathcal{C}}_{t_-}, \widehat{q}^{\mathcal{C}}_{t_-})$ by $\varepsilon/16$. The actual algorithm starts from ν_{t_+} rather than $q^{\mathcal{C}}_{t_+}$. Since both initial distributions are propagated by the same approximate reverse kernels, we have

$$\mathsf{TV}\Big(\widehat{q}^{\mathcal{C}}_{t_-}, \mathrm{law}(Y_0 \mid \mathcal{C})\Big) \leqslant \mathsf{TV}(q^{\mathcal{C}}_{t_+}, \nu_{t_+}) \leqslant \varepsilon/32\,.$$

Combining this with the preceding endpoint bound, with triangle inequality, gives

$$\mathsf{TV}\Big(q^{\mathcal{C}}_{t_-}, \mathrm{law}(Y_0 \mid \mathcal{C})\Big) \leqslant \varepsilon/16 + \varepsilon/32 = 3\varepsilon/32\,.$$

If the final observation had law $q^{\mathcal{C}}_{t_-}$, Lemma G.8 would give

$$\mathop{\mathbb{P}}_{q^{\mathcal{C}}_{t_-}}[\|\widetilde{m}_{t_-} - m_{t_-}\|_2 > 1/4] \leqslant Ce^{-2B\Lambda} \leqslant \varepsilon/32\,.$$

By Lemma G.15 and Eq. (95), its rounded output law would then be within $\varepsilon/16$ of $q^{\mathcal{C}}$ in total variation. Rounding the actual endpoint adds at most $\mathsf{TV}\Big(q^{\mathcal{C}}_{t_-}, \mathrm{law}(Y_0 \mid \mathcal{C})\Big)$ of total variation error. Thus, conditional on a successful window search, the total error is at most $3\varepsilon/32 + \varepsilon/16 = 5\varepsilon/32$. Averaging over the window search adds at most its failure probability $\varepsilon/8$, using $\mathsf{TV} \leqslant 1$ on the failure event, so the output law is within $5\varepsilon/32 + \varepsilon/8 = 9\varepsilon/32 \leqslant \varepsilon$ of $q^{\mathcal{C}}$ in total variation. The algorithm fails with probability at most $\delta/2 + \delta/4 < \delta$ over the choice of $\mathcal{C}$, completing the proof of correctness.

The search uses $O(\Lambda/w) = O(\sqrt{d\Lambda})$ queries. The reverse steps use $N = O\left(\frac{dw+dw^2}{\varepsilon^2} + 1\right) = O\left(\frac{\sqrt{d\Lambda}+\Lambda}{\varepsilon^2} + 1\right)$ queries, and rounding uses one more query, completing the proof. □

H $\widetilde{\Omega}(d)$ LOWER BOUND FOR MASKED DIFFUSION

We now consider the same random empirical measure under masked diffusion and exhibit an approximate score oracle for which any sampler requires $\widetilde{\Omega}(d)$ queries, in contrast to uniform and Gaussian diffusion. The intuition for this lower bound, which we state formally in Theorem H.1 below, is as follows. If the codebook defining the empirical measure has size $|\mathcal{C}| = 2^k$, and a total of m coordinates have been revealed, a fixed partial assignment agrees with about 2^{k-m} points of $\mathcal{C}$. When m is well below k, these points have nearly balanced signs in the remaining coordinates. When m is well above k, a partial assignment chosen without knowledge of $\mathcal{C}$ is unlikely to agree with any point at all. We will use these two facts to construct an accurate masked diffusion oracle that usually returns uniform marginals in both ranges. Only queries at values of m close to the unknown value k need reveal information about the target.

Let $\mathcal{K}^{\mathsf{M}}_d := \{k \in \mathbb{Z} : \kappa_- \leqslant k \log 2/d \leqslant \kappa_+\}$, and consider the prior under which $k \sim \mathrm{Unif}(\mathcal{K}^{\mathsf{M}}_d)$ and, given k, $\mathcal{C}$ is a uniformly random 2^k-element subset of $\mathcal{X}$. For a partial assignment (I, x_I),

write $\mathsf{match}_I(x_I) \coloneqq \{y \in \mathcal{C} : y_I = x_I\}$ for the set of codewords matching it. Then (I, x_I) is consistent with $q^{\mathcal{C}}$ if and only if $\mathsf{match}_I(x_I) \neq \varnothing$, in which case

$$q^{\mathcal{C}}_{i|I}(a \mid x_I) = \frac{|\{y \in \mathsf{match}_I(x_I) : y_i = a\}|}{|\mathsf{match}_I(x_I)|}, \qquad i \notin I\,.$$

We measure the accuracy of an oracle $\mathcal{O}$ by its level-m score error $\varepsilon_{\mathsf{mask}}(\mathcal{O}; q^{\mathcal{C}}, m)$ from Definition B.5, in the form given by Lemma B.6. Recall that this constrains $\mathcal{O}$ only on consistent partial assignments; on inconsistent partial assignments, the oracle can return arbitrary values.

Theorem H.1 (Masked diffusion lower bound)**.** *Fix $\rho \in (0, 1/4)$. There are constants $c, C, d_0 > 0$, depending only on $[\kappa_-, \kappa_+]$ and ρ, such that the following holds for $d \geq d_0$. Given $\eta \in (0, 1/10)$, let*

$$\ell = \left\lceil \log_2 \frac{Cd}{\eta\rho^2} \right\rceil . \tag{96}$$

There is a family of oracles $\{\mathcal{O}_{k,\mathcal{C}}\}_{k \in \mathcal{K}^{\mathsf{M}}_d, |\mathcal{C}| = 2^k}$ satisfying

$$\max_{0 \leqslant m < d} \varepsilon_{\mathsf{mask}}(\mathcal{O}_{k,\mathcal{C}}; q^{\mathcal{C}}, m) \leqslant \eta, \qquad k \in \mathcal{K}^{\mathsf{M}}_d, |\mathcal{C}| = 2^k\,, \tag{97}$$

such that every adaptive randomized algorithm $\mathcal{A}$ making at most

$$Q \leqslant c\rho^2 \frac{d}{\ell}$$

queries, and which does not have knowledge of $k, \mathcal{C}$, has output law $\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}$ satisfying

$$\mathop{\mathbb{P}}_{\substack{k \sim \mathrm{Unif}(\mathcal{K}^{\mathsf{M}}_d) \\ \mathcal{C} \sim \mathrm{Unif}\left(\binom{\{\pm 1\}^d}{2^k}\right)}} [\mathsf{TV}(\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}, q^{\mathcal{C}}) \geq 1 - \rho] \geq 1 - \rho\,.$$

In particular, for constant ρ and any polynomially small $\eta = d^{-\Theta(1)}$, sampling with at most constant error for at least a constant fraction of the codebooks requires $\Omega(d/\log d)$ queries.

In Section H.1 we describe the oracle construction. In Section H.2, we prove the lower bound.

H.1 Adversarial oracle construction

Define the oracle $\mathcal{O}_0$ to return $\mathrm{Unif}(\{\pm 1\})$ in every unrevealed coordinate. We first check its average error when few coordinates are revealed.

Lemma H.2. *For every $k \in \mathcal{K}^{\mathsf{M}}_d$ and $0 \leqslant m < d$,*

$$\mathbb{E}_{\mathcal{C}} \varepsilon_{\mathsf{mask}}(\mathcal{O}_0; q^{\mathcal{C}}, m) \leqslant 2^{m-k}\,.$$

Proof. Fix I with $|I| = m$ and $i \notin I$. For a fixed assignment x_I, define $N^{\mathcal{C}}_I(x_I) = |\mathsf{match}_I(x_I)|$. Conditional on $N^{\mathcal{C}}_I(x_I) > 0$, these $N^{\mathcal{C}}_I(x_I)$ points form a uniformly drawn subset of the subcube $\{y : y_I = x_I\}$. Exactly half of this subcube has $y_i = +1$. Define $P^{\mathcal{C}}_{i|I}(x_I) \coloneqq |\{y \in \mathsf{match}_I(x_I) : y_i = +1\}|/N^{\mathcal{C}}_I(x_I)$ as the fraction of points in $\mathsf{match}_I(x_I)$ with $y_i = +1$, so when $N^{\mathcal{C}}_I(x_I) > 0$,

$$\mathbb{E}_{\mathcal{C}}[P^{\mathcal{C}}_{i|I}(x_I) \mid N^{\mathcal{C}}_I(x_I)] = \frac{1}{2}, \qquad \mathbb{E}_{\mathcal{C}}\left[\left(P^{\mathcal{C}}_{i|I}(x_I) - \frac{1}{2}\right)^2 \mid N^{\mathcal{C}}_I(x_I)\right] \leqslant \frac{1}{4N^{\mathcal{C}}_I(x_I)}\,.$$

The variance bound follows since we are sampling $\mathsf{match}_I(x_I)$ without replacement. The inequality $\log u \leqslant u - 1$ also gives

$$\mathsf{KL}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(1/2)) \leqslant 4(p - 1/2)^2\,.$$

Thus for any $N^{\mathcal{C}}_I(x_I) > 0$,

$$\mathbb{E}_{\mathcal{C}}\left[N^{\mathcal{C}}_I(x_I)\, \mathsf{KL}(\mathrm{Ber}(P^{\mathcal{C}}_{i|I}(x_I)) \,\|\, \mathrm{Ber}(1/2)) \mid N^{\mathcal{C}}_I(x_I)\right] \leqslant 4N^{\mathcal{C}}_I(x_I)\, \mathbb{E}_{\mathcal{C}}\left[\left(P^{\mathcal{C}}_{i|I}(x_I) - \frac{1}{2}\right)^2 \mid N^{\mathcal{C}}_I(x_I)\right] \leqslant 1\,.$$

When $N_I^{\mathcal{C}}(x_I) = 0$, define $N_I^{\mathcal{C}}(x_I)\,\mathsf{KL}(\mathrm{Ber}(P_{i|I}^{\mathcal{C}}(x_I)) \,\|\, \mathrm{Ber}(1/2)) = 0$.

Now fix $\mathcal{C}$. A sample $X_0 \sim q^{\mathcal{C}}$ satisfies $X_{0,I} = x_I$ with probability $N_I^{\mathcal{C}}(x_I)/2^k$. Hence the expected KL error is

$$\mathbb{E}_{X_0 \sim q^{\mathcal{C}}} \mathsf{KL}(q_{i|I}^{\mathcal{C}}(\cdot \mid X_{0,I}) \,\|\, \mathrm{Ber}(1/2)) = \frac{1}{2^k} \sum_{x_I \in \{\pm 1\}^I} N_I^{\mathcal{C}}(x_I)\,\mathsf{KL}(\mathrm{Ber}(P_{i|I}^{\mathcal{C}}(x_I)) \,\|\, \mathrm{Ber}(1/2))\,.$$

Now average over $\mathcal{C}$. Each of the 2^m summands has expectation at most one by the preceding calculation, so the result is at most 2^{m-k}. Averaging over I and i proves the lemma. □

Call $\mathcal{C}$ *good* if the answers provided by $\mathcal{O}_0$ satisfy the desired error bound at every level $m \leqslant k - \ell$:

$$\varepsilon_{\mathsf{mask}}(\mathcal{O}_0; q^{\mathcal{C}}, m) \leqslant \eta \quad \text{for every } 0 \leqslant m \leqslant k - \ell\,. \tag{98}$$

Lemma H.2, Markov's inequality, and a union bound over the levels $0 \leqslant m \leqslant k - \ell$ imply

$$\mathop{\mathbb{P}}_{\mathcal{C}}[\mathcal{C} \text{ is not good}] \leqslant \frac{1}{\eta} \sum_{m=0}^{k-\ell} 2^{m-k} \leqslant \frac{2^{1-\ell}}{\eta}\,. \tag{99}$$

Definition H.3 (Adversarial masked diffusion oracle)**.** *Define the oracle $\mathcal{O}_{k,\mathcal{C}}$ as follows:*

- *For a good $\mathcal{C}$, it returns uniform marginals at all levels $m \leqslant k - \ell$. At levels $m > k - \ell$, it returns the exact conditional marginals on consistent assignments and uniform marginals on inconsistent ones.*
- *For $\mathcal{C}$ that is not good, at all levels m, it returns the exact conditional marginals on every consistent assignment and uniform marginals on inconsistent ones.*

Proposition H.4 (Validity of the oracle)**.** *For every $k \in \mathcal{K}_d^{\mathsf{M}}$ and $|\mathcal{C}| = 2^k$, the oracle $\mathcal{O}_{k,\mathcal{C}}$ satisfies $\max_{0 \leqslant m < d} \varepsilon_{\mathsf{mask}}(\mathcal{O}_{k,\mathcal{C}}; q^{\mathcal{C}}, m) \leqslant \eta$, i.e., Eq. (97).*

Proof. By Lemma B.6, $\varepsilon_{\mathsf{mask}}(\mathcal{O}_{k,\mathcal{C}}; q^{\mathcal{C}}, m)$ depends only on the answers of $\mathcal{O}_{k,\mathcal{C}}$ on consistent assignments with m revealed coordinates. These answers are exact, so the error at level m vanishes, except when $\mathcal{C}$ is good and $m \leqslant k - \ell$, in which case $\mathcal{O}_{k,\mathcal{C}}$ agrees with $\mathcal{O}_0$ and the error is at most η by Eq. (98). □

H.2 Distribution indistinguishability: the lower bound proof

Let R denote all the internal randomness of an algorithm $\mathcal{A}$. The following compares the algorithm's interaction with the constructed oracle to its interaction with $\mathcal{O}_0$. For the latter, every answer is uniform, so the entire query sequence is fully determined by the internal randomness of the algorithm R.

Proposition H.5 (Coupling argument)**.** *Couple an algorithm's run against the oracle $\mathcal{O}_{k,\mathcal{C}}$ with its run against the oracle $\mathcal{O}_0$, using the same internal randomness R. For an algorithm making at most Q queries, let $p_{\mathrm{diff}}(k, \mathcal{C})$ be the conditional probability over R given $k, \mathcal{C}$ that the runs receive different oracle answers at some point. Then*

$$\mathbb{E}_{k,\mathcal{C}}\, p_{\mathrm{diff}}(k, \mathcal{C}) \leqslant \frac{(2\ell + 1)Q}{|\mathcal{K}_d^{\mathsf{M}}|} + \frac{2^{1-\ell}}{\eta} + Q2^{-\ell}\,. \tag{100}$$

Proof. Let $\mathcal{D}$ denote the event that the two coupled runs receive different oracle answers at some point. By definition,

$$\mathbb{E}_{k,\mathcal{C}}\, p_{\mathrm{diff}}(k, \mathcal{C}) = \mathbb{E}_{k,\mathcal{C}} \mathop{\mathbb{P}}_{R}[\mathcal{D} \mid k, \mathcal{C}] = \mathop{\mathbb{P}}_{k,\mathcal{C},R}[\mathcal{D}] = \mathbb{E}_R \mathop{\mathbb{P}}_{k,\mathcal{C}}[\mathcal{D} \mid R]\,. \tag{101}$$

Now condition on any fixed realization of the internal randomness $R = r$. We will directly bound $\mathbb{P}_{k,\mathcal{C}}[\mathcal{D} \mid R = r]$ uniformly over r.

It is enough to bound the probability of a first different answer, because the queries agree up to that point.

Run the algorithm against oracle $\mathcal{O}_0$. This fixes queries $(I_1, x_{I_1}), \dots, (I_Q, x_{I_Q})$, padding with irrelevant queries if the algorithm stops early. The sequence is independent of both k and $\mathcal{C}$. Consider the randomness of k. For any query (I_j, x_{I_j}), $m_j := |I_j|$ is within ℓ of at most $2\ell + 1$ possible integers k. A union bound shows that there exists a query with $|m_j - k| \leqslant \ell$ with probability at most $(2\ell + 1)Q/|\mathcal{K}_d^{\mathsf{M}}|$ over k.

Fix a k outside all these intervals $\{[m_j - \ell, m_j + \ell]\}_{j=1}^{Q}$. We also remove the event that $\mathcal{C}$ is not good, whose probability is bounded by Eq. (99).

High-noise regime ($m < k - \ell$). At any query (I, x_I) with $m < k - \ell$, Definition H.3 returns uniform marginals when $\mathcal{C}$ is good, regardless of whether the assignment is consistent. Its answer agrees with $\mathcal{O}_0$.

Low-noise regime ($m > k+\ell$). Fix one query (I, x_I) with $m > k+\ell$. The subcube $\{y : y_I = x_I\}$ specified by this assignment has 2^{d-m} points, so

$$\mathop{\mathbb{P}}_{\mathcal{C}}[\mathsf{match}_I(x_I) \neq \varnothing] \leqslant \mathbb{E}_{\mathcal{C}}|\mathsf{match}_I(x_I)| = 2^{k-m} \leqslant 2^{-\ell}\,.$$

This probability is under the original uniform choice of $\mathcal{C}$ (we do not condition on $\mathcal{C}$ being good). A union bound over the at most Q queries with $m_j > k + \ell$ therefore shows that the probability that any such queried assignment is consistent is at most $Q2^{-\ell}$.

Thus, for the fixed realization of $R = r$, if no query size m_j lies in $[k - \ell, k + \ell]$, if $\mathcal{C}$ is good, and if every queried assignment with $m_j > k + \ell$ is inconsistent with $\mathcal{C}$, then every query in the run against $\mathcal{O}_{k,\mathcal{C}}$ receives exactly the same uniform answer as in the run against $\mathcal{O}_0$. Indeed, before a first different answer the two runs have identical queries and answers, and hence make the same next query. By induction over the queries, no first different answer can occur under these three conditions.

Consequently, uniformly over $R = r$,

$$\mathop{\mathbb{P}}_{k,\mathcal{C}}[\mathcal{D} \mid R = r] \leqslant \frac{(2\ell + 1)Q}{|\mathcal{K}_d^{\mathsf{M}}|} + \frac{2^{1-\ell}}{\eta} + Q2^{-\ell}\,. \tag{102}$$

Averaging Eq. (102) over R and using Eq. (101) gives Eq. (100). □

We can now complete the proof of the main result of this section:

Proof of Theorem H.1. By Proposition H.4, the oracle $\mathcal{O}_{k,\mathcal{C}}$ satisfies the accuracy requirement for every instance. Fix any algorithm with query budget Q, and let ν be its output law when every oracle answer is uniform. This law is independent of $k, \mathcal{C}$.

For each fixed instance $(k, \mathcal{C})$, couple the two runs of the algorithm, one against $\mathcal{O}_{k,\mathcal{C}}$ and the other against $\mathcal{O}_0$, by using the same internal randomness R, as in Proposition H.5. Let $Y_{k,\mathcal{C}}(R)$ and $Y_0(R)$ denote the corresponding outputs. Their marginal laws over R are respectively $\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}$ and ν. Hence, by the coupling inequality,

$$\begin{aligned}
\mathsf{TV}(\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}, \nu) &\leqslant \mathop{\mathbb{P}}_{R}[Y_{k,\mathcal{C}}(R) \neq Y_0(R)] \\
&\leqslant \mathop{\mathbb{P}}_{R}[\text{the two runs receive different oracle answers}] \\
&= p_{\text{diff}}(k, \mathcal{C})\,.
\end{aligned}$$

The second inequality holds because, for every fixed realization of R, identical oracle answers yield the same final output.

Since $q^{\mathcal{C}}$ is supported on $\mathcal{C}$, it follows that

$$\mathsf{TV}(\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}, q^{\mathcal{C}}) \geq 1 - \widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}(\mathcal{C}) \geq 1 - \nu(\mathcal{C}) - \mathsf{TV}(\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}, \nu) \geq 1 - \nu(\mathcal{C}) - p_{\text{diff}}(k, \mathcal{C})\,. \tag{103}$$

A fixed point belongs to a uniform 2^k-element subset with probability 2^{k-d}. Averaging according to any output law ν independent of $\mathcal{C}$ thus gives

$$\mathbb{E}_{\mathcal{C}}\nu(\mathcal{C}) = \mathbb{E}_{X\sim\nu}\mathbb{E}_{\mathcal{C}}\,\mathbb{1}[X \in \mathcal{C}] = 2^{k-d} \leqslant e^{-(\log 2 - \kappa_+)d}\,.$$

We have $|\mathcal{K}_d^{\mathsf{M}}| \geq c_1 d$ for some constant $c_1 > 0$ depending only on $[\kappa_-, \kappa_+]$, for all sufficiently large d. Choose the constant c in the query budget small enough that the first term of Eq. (100) is at most $\rho^2/4$. Also $Q \leqslant d$ for this choice. Taking C in Eq. (96) sufficiently large bounds the second and third terms in Eq. (100) by $\rho^2/4$ in total. For sufficiently large d, the expected mass $\mathbb{E}_{\mathcal{C}}\nu(\mathcal{C})$ is at most $\rho^2/4$. Hence,

$$\mathbb{E}_{k,\mathcal{C}}[\nu(\mathcal{C}) + p_{\mathrm{diff}}(k, \mathcal{C})] \leqslant \rho^2 .$$

Markov's inequality shows that $\mathbb{P}_{k,\mathcal{C}}[\nu(\mathcal{C}) + p_{\mathrm{diff}}(k, \mathcal{C}) \leqslant \rho] \geq 1 - \rho$. Eq. (103) proves the theorem. Finally, $\ell = O(\log d)$ for polynomially small η and constant ρ, giving $\Omega(d/\log d)$ queries. □

Note that implicit in the argument above was that the critical window for masked diffusion has width which is only an $O(\log d/d)$ (rather than $O(\sqrt{\log d/d})$ as in Gaussian and uniform diffusion) fraction of the total range of noise levels. We make this precise with the following analogue of Propositions F.4 and G.4: above the window, the revealed coordinates of a random codeword are close to uniform, while below it, the original point can be recovered with high probability. In analogy with $\mathrm{err}_{\mathcal{C}}^{\mathsf{U}}(t)$ and $\mathrm{err}_{\mathcal{C}}^{\mathsf{G}}(t)$, define

$$\mathrm{err}_{\mathcal{C}}^{\mathsf{M}}(m) := \inf_{\widehat{z}(\cdot)} \mathop{\mathbb{P}}_{z \sim q^{\mathcal{C}},\, I \sim \mathrm{Unif}\left(\binom{[d]}{m}\right)} [\widehat{z}(I, z_I) \neq z] ,$$

the smallest error of any decoder that observes m uniformly random coordinates of a sample from $q^{\mathcal{C}}$.

Proposition H.6 (Critical window). *Let $|\mathcal{C}| = 2^k$ and $\ell \geq 1$, and fix $0 \leqslant m < d$. If $m \leqslant k - \ell$, then for every $I \subseteq [d]$ with $|I| = m$ and $X_0 \sim q^{\mathcal{C}}$ (with $\mathrm{law}(X_{0,I})$ taken conditionally on $\mathcal{C}$),*

$$\mathbb{E}_{\mathcal{C}}\, \mathsf{TV}\big(\mathrm{law}(X_{0,I}), \mathrm{Unif}(\{\pm 1\}^I)\big) \leqslant 2^{-\ell/2} . \tag{104}$$

If $m \geq k + \ell$, then

$$\mathbb{E}_{\mathcal{C}}\, \mathrm{err}_{\mathcal{C}}^{\mathsf{M}}(m) \leqslant 2^{-\ell} . \tag{105}$$

Proof. Fix I with $|I| = m$ and write $N(x_I) := |\mathsf{match}_I(x_I)|$, so that $\mathrm{law}(X_{0,I}) = N(\cdot)/2^k$. Since $\mathcal{C}$ is a uniformly random 2^k-element subset of $\mathcal{X}$, each $N(x_I)$ is hypergeometric with mean 2^{k-m} and variance at most 2^{k-m}. Thus

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\, \mathsf{TV}\big(\mathrm{law}(X_{0,I}), \mathrm{Unif}(\{\pm 1\}^I)\big) &= \frac{1}{2} \sum_{x_I \in \{\pm 1\}^I} \mathbb{E}_{\mathcal{C}} \big| N(x_I) 2^{-k} - 2^{-m} \big| \\
&\leqslant \frac{1}{2} \cdot 2^m \cdot 2^{-k} \sqrt{2^{k-m}} = \frac{1}{2} \cdot 2^{-(k-m)/2} \leqslant 2^{-\ell/2} \qquad \text{for } m \leqslant k - \ell ,
\end{aligned}$$

which gives Eq. (104). For Eq. (105), consider the decoder which outputs the unique element of $\mathsf{match}_I(z_I)$ whenever this set is a singleton, and something arbitrary otherwise. It is incorrect only if some other codeword agrees with z on I. Conditionally on z, the remaining $2^k - 1$ codewords form a uniformly random subset of $\mathcal{X} \setminus \{z\}$, and each agrees with z on I with probability at most 2^{-m}, so by a union bound $\mathbb{E}_{\mathcal{C}}\, \mathrm{err}_{\mathcal{C}}^{\mathsf{M}}(m) \leqslant 2^{k-m} \leqslant 2^{-\ell}$ for $m \geq k + \ell$. □

I $\widetilde{\Omega}(\sqrt{d})$ LOWER BOUND FOR UNIFORM DIFFUSION

The upper bound in Section F searches for the critical time $t^\star_{\mathsf{unif}}$ defined in Eq. (22). We show that this search cannot be avoided with approximate scores. The hard distribution family contains $\widetilde{\Omega}(\sqrt{d})$ possible values of $t^\star_{\mathsf{unif}}$, separated by more than the width $w = \widetilde{\Theta}(1/\sqrt{d})$ of the critical window defined in Eq. (107) below, and a query far from the realized value usually receives the score of the uniform distribution.

This is the continuous-time counterpart of the masked lower bound in Section H. There, the hidden transition occurs when the number of revealed coordinates passes the hidden integer k. Here it occurs when the query time passes $t^\star_{\mathsf{unif}}$. Above $t^\star_{\mathsf{unif}}$, the noised random empirical measure is close to uniform. Below $t^\star_{\mathsf{unif}}$, a genuine sample is close to a point of $\mathcal{C}$, but a fixed query chosen without knowing $\mathcal{C}$ is unlikely to be close to any point of $\mathcal{C}$.

For every integer k satisfying $\kappa_- \leqslant k \log 2/d \leqslant \kappa_+$, let $t^\star_{\mathsf{unif}}(k)$ be the unique solution of

$$d\, \mathsf{I}_{\mathsf{unif}}(t^\star_{\mathsf{unif}}(k)) = k \log 2 = \log M\,. \tag{106}$$

The derivative

$$\mathsf{I}'_{\mathsf{unif}}(t) = -\frac{e^{-t}}{2} \log \frac{1+e^{-t}}{1-e^{-t}}$$

is bounded above and below in magnitude on $[\tau_-, \tau_+]$.

Given $A, C_\Lambda > 0$, $\eta \in (0, 1/10)$, and $\rho \in (0, 1/4)$, define

$$\Lambda = \log \frac{C_\Lambda d}{\eta \rho^2}\,, \qquad w = A\sqrt{\frac{\Lambda}{d}}\,. \tag{107}$$

Fix a sufficiently large constant $D > 0$. Starting with the smallest admissible integer k, retain every $\lceil Ddw \rceil$-th one, and denote the resulting set by

$$\mathcal{K}_d^{\mathsf{U}} := \left\{ \left\lceil \frac{\kappa_- d}{\log 2} \right\rceil + j\lceil Ddw \rceil : j = 0, 1, \dots, \left\lfloor \frac{\kappa_+ d/\log 2 - \lceil \kappa_- d/\log 2 \rceil}{\lceil Ddw \rceil} \right\rfloor \right\}\,.$$

If $k < k'$ are consecutive elements of $\mathcal{K}_d^{\mathsf{U}}$, the mean value theorem applied to Eq. (106) gives, for some $\xi \in (t^\star_{\mathsf{unif}}(k'), t^\star_{\mathsf{unif}}(k))$,

$$t^\star_{\mathsf{unif}}(k) - t^\star_{\mathsf{unif}}(k') = \frac{(k'-k)\log 2/d}{-\mathsf{I}'_{\mathsf{unif}}(\xi)}\,.$$

Consequently, after choosing D large enough,

$$t^\star_{\mathsf{unif}}(k) - t^\star_{\mathsf{unif}}(k') \geq 10w, \qquad |\mathcal{K}_d^{\mathsf{U}}| \geq \frac{cd}{\lceil Ddw \rceil} \geq \frac{c_{\mathsf{U}}}{w}\,, \tag{108}$$

for all sufficiently large d. In particular,

$$|\mathcal{K}_d^{\mathsf{U}}| \geq \frac{c_{\mathsf{U}}}{A}\sqrt{\frac{d}{\Lambda}}\,. \tag{109}$$

We choose k uniformly from $\mathcal{K}_d^{\mathsf{U}}$ and, conditional on k, choose $\mathcal{C}$ uniformly among the 2^k-element subsets of $\mathcal{X}$. Once k is fixed, we write $t^\star_{\mathsf{unif}}$ for $t^\star_{\mathsf{unif}}(k)$.

Throughout, $s_t^{\mathcal{C}}$ denotes the exact score of $q_t^{\mathcal{C}}$, that is, $s_t^{\mathcal{C}}(x)[i,a] = q_t^{\mathcal{C}}(x^{i\leftarrow a})/q_t^{\mathcal{C}}(x)$. We measure the accuracy of a score oracle $\mathcal{O}$ by $\varepsilon_{\mathsf{unif}}(\mathcal{O}; q^{\mathcal{C}}, t)$ from Definition 2.1.

Theorem I.1 (Uniform diffusion lower bound)**.** *Fix $\rho \in (0, 1/4)$. There are constants $A, C_\Lambda, c, c_0, d_0 > 0$, depending only on $[\kappa_-, \kappa_+]$ and ρ, such that the following holds for $d \geq d_0$. Given $\eta \in (0, 1/10)$, take Λ, w, and $\mathcal{K}_d^{\mathsf{U}}$ as defined above, and suppose $\Lambda \leqslant c_0 d$. There is a family of score oracles $\{\mathcal{O}_{k,\mathcal{C}}\}_{k \in \mathcal{K}_d^{\mathsf{U}}, |\mathcal{C}| = 2^k}$, each returning normalized score vectors with strictly positive entries, satisfying*

$$\sup_{t>0} \varepsilon_{\mathsf{unif}}(\mathcal{O}_{k,\mathcal{C}}; q^{\mathcal{C}}, t) \leqslant \eta, \qquad k \in \mathcal{K}_d^{\mathsf{U}}, \quad |\mathcal{C}| = 2^k\,, \tag{110}$$

such that every adaptive randomized algorithm $\mathcal{A}$ that is not given k or $\mathcal{C}$ and makes at most

$$Q \leqslant c\rho^2 \sqrt{\frac{d}{\Lambda}} \tag{111}$$

score oracle queries has output law $\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}$ satisfying

$$\mathop{\mathbb{P}}_{\substack{k \sim \mathrm{Unif}(\mathcal{K}_d^{\mathsf{U}}) \\ \mathcal{C} \sim \mathrm{Unif}\left(\binom{\mathcal{X}}{2^k}\right)}} [\mathsf{TV}(\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}, q^{\mathcal{C}}) \geq 1 - \rho] \geq 1 - \rho\,. \tag{112}$$

In particular, for constant ρ and any polynomially small $\eta = d^{-\Theta(1)}$, sampling with at most constant error for at least a constant fraction of the instances requires $\Omega(\sqrt{d/\log d})$ score oracle queries.

For the rest of the appendix, Λ and w have the values in Eq. (107). We may increase A, C_Λ, d_0 and decrease c_0 later when necessary so that $\log d \leqslant \Lambda \leqslant c_0 d$ and $w \leqslant \min\{\tau_-/8, 1/8\}$. Fix a constant $\gamma > 0$ sufficiently small in terms of $[\tau_-, \tau_+]$.

In Section I.1, we construct the oracle family and prove its accuracy. In Section I.2, we prove the query lower bound.

I.1 Adversarial oracle construction

Let $\mathbf{1} \in \mathbb{R}^{[d]\times\{-1,1\}}$ be the score vector defined by $\mathbf{1}[i,a] = 1$ for every $i \in [d]$ and $a \in \{-1,1\}$, and define the reference oracle by $\mathcal{O}_0(t,x) := \mathbf{1}$ for every query (t,x). This is the exact score oracle of $\mathrm{Unif}(\mathcal{X})$. Proposition F.4 already shows that $q_t^{\mathcal{C}}$ is close to $\mathrm{Unif}(\mathcal{X})$ above $t^\star_{\mathsf{unif}}$; the following lemma records the corresponding statement for the score error.

Lemma I.2. *For every $K > 0$, there are constants $A_K, c_K, d_K > 0$, depending only on $[\kappa_-, \kappa_+]$ and K, such that, whenever $A \geq A_K$, $d \geq d_K$, $\log d \leqslant \Lambda \leqslant c_K d$, and $t \geq t^\star_{\mathsf{unif}}(k) + 4w$,*

$$\mathbb{E}_{\mathcal{C}}\, \mathsf{TV}(q_t^{\mathcal{C}}, \mathrm{Unif}(\mathcal{X})) \leqslant e^{-K\Lambda}, \qquad \mathbb{E}_{\mathcal{C}}\, \varepsilon_{\mathsf{unif}}(\mathcal{O}_0; q^{\mathcal{C}}, t) \leqslant e^{-K\Lambda}\,.$$

Both inequalities hold uniformly over $k \in \mathcal{K}_d^{\mathsf{U}}$ and $t \geq t^\star_{\mathsf{unif}}(k) + 4w$.

Proof. Since $t \geq t^\star_{\mathsf{unif}} + 4w$, Proposition F.4 with exponent $K+2$, whose constants at that exponent we take as the A_K, c_K, d_K of the present lemma, gives

$$\mathbb{E}_{\mathcal{C}}\, \mathsf{TV}(q_t^{\mathcal{C}}, \mathrm{Unif}(\mathcal{X})) \leqslant e^{-(K+2)\Lambda} \leqslant e^{-K\Lambda}\,.$$

Eq. (32) places the exact scores in a fixed compact subinterval of $(0,\infty)$, on which $\psi(v,1) \leqslant C|v-1|$. For each $\mathcal{C}$ and each coordinate i, we then have

$$\begin{aligned}\mathbb{E}_{X\sim q_t^{\mathcal{C}}} \psi(s_t^{\mathcal{C}}(X)[i,-X_i],1) &\leqslant C \sum_{x\in\mathcal{X}} \left|q_t^{\mathcal{C}}(x^{i\leftarrow -x_i}) - q_t^{\mathcal{C}}(x)\right| \\ &\leqslant 4C\, \mathsf{TV}(q_t^{\mathcal{C}}, \mathrm{Unif}(\mathcal{X}))\,. \end{aligned} \tag{113}$$

The second inequality follows by inserting 2^{-d} and using the fact that a coordinate flip permutes $\mathcal{X}$. Summing Eq. (113) over i and averaging over $\mathcal{C}$ gives

$$\mathbb{E}_{\mathcal{C}}\, \varepsilon_{\mathsf{unif}}(\mathcal{O}_0; q^{\mathcal{C}}, t) \leqslant 4Cd\, e^{-(K+2)\Lambda} \leqslant e^{-K\Lambda}\,,$$

where the last inequality holds after increasing d_K. □

Well below $t^\star_{\mathsf{unif}}$, a sample X_t from $q_t^{\mathcal{C}}$ is usually close to the point X_0 of $\mathcal{C}$ from which it was generated. For $t > 0$, define

$$R_t = \lceil d(\beta_t + \gamma w)\rceil + 1, \qquad \mathcal{N}_{\mathcal{C},t} = \left\{ x \in \mathcal{X} : \min_{y\in\mathcal{C}} d_H(x,y) \leqslant R_t \right\}.$$

The oracle will return the exact score on the neighborhood $\mathcal{N}_{\mathcal{C},t}$ and the all-one score outside it. The next lemma separates the two facts needed later: a genuine sample from $q_t^{\mathcal{C}}$ usually lies in this neighborhood, whereas a fixed query usually does not.

Lemma I.3. *There are constants $c, C, c_1, d_1 > 0$, depending only on $[\kappa_-, \kappa_+]$, such that, whenever $d \geq d_1$, $\log d \leqslant dw^2 \leqslant c_1 d$, and $0 < t \leqslant t^\star_{\mathsf{unif}} - 4w$, the following statements hold uniformly over the allowed values of k and t.*

(i) For every $\mathcal{C}$,

$$q_t^{\mathcal{C}}(\mathcal{N}_{\mathcal{C},t}^{c}) \leqslant e^{-2\gamma^2 dw^2}\,. \tag{114}$$

If the score output of the oracle $\mathcal{O}$ equals $s_t^{\mathcal{C}}$ on $\mathcal{N}_{\mathcal{C},t}$ and the all-one score on $\mathcal{N}_{\mathcal{C},t}^{c}$, then

$$\varepsilon_{\mathsf{unif}}(\mathcal{O}; q^{\mathcal{C}}, t) \leqslant Cd(1+\log d)e^{-cdw^2}\,. \tag{115}$$

(ii) For every fixed $x \in \mathcal{X}$,

$$\mathop{\mathbb{P}}_{\mathcal{C}}[x \in \mathcal{N}_{\mathcal{C},t}] \leqslant e^{-cdw}\,. \tag{116}$$

Proof. Fix k and write $t^\star_{\mathsf{unif}} = t^\star_{\mathsf{unif}}(k)$. Since $\beta'_r = e^{-r}/2$ and $0 \leqslant t < t^\star_{\mathsf{unif}} \leqslant \tau_+$,

$$\beta_{t^\star_{\mathsf{unif}}} - \beta_t = \int_t^{t^\star_{\mathsf{unif}}} \frac{e^{-r}}{2}\,\mathrm{d}r \geq \frac{e^{-\tau_+}}{2}(t^\star_{\mathsf{unif}} - t) \geq 2e^{-\tau_+}w\,.$$

Choose $\gamma < e^{-\tau_+}/2$. Since $dw^2 \geq \log d$, i.e., $w \geq \sqrt{\log d/d}$, after increasing d_1 we have

$$\frac{R_t}{d} \leqslant \beta_t + \gamma w + \frac{2}{d} \leqslant \beta_{t^\star_{\mathsf{unif}}} - cw < \frac{1}{2}\,. \tag{117}$$

We start by showing Eq. (114). Fix $\mathcal{C}$, draw $Y \sim q^{\mathcal{C}}$, and generate $X \sim (K_t^{\mathsf{U}})^{\otimes d}(\cdot \mid Y)$. Then $d_H(X,Y) \sim \mathrm{Bin}(d,\beta_t)$ and $\min_{z\in\mathcal{C}} d_H(X,z) \leqslant d_H(X,Y)$. By Hoeffding's inequality, we have

$$q_t^{\mathcal{C}}(\mathcal{N}_{\mathcal{C},t}^c) \leqslant \mathbb{P}[\mathrm{Bin}(d,\beta_t) \geq R_t] \leqslant e^{-2\gamma^2 dw^2}\,,$$

proving Eq. (114).

We next bound the score error outside $\mathcal{N}_{\mathcal{C},t}$. For every $x \in \mathcal{X}$ and $i \in [d]$, the binary likelihood ratio in Lemma B.2 gives

$$\frac{\beta_t}{1-\beta_t} \leqslant s_t^{\mathcal{C}}(x)[i,-x_i] \leqslant \frac{1-\beta_t}{\beta_t}\,.$$

Hence

$$q_t^{\mathcal{C}}(x)\psi(s_t^{\mathcal{C}}(x)[i,-x_i],1) \leqslant q_t^{\mathcal{C}}(x) + \log\frac{1-\beta_t}{\beta_t}\, q_t^{\mathcal{C}}(x^{i\leftarrow -x_i})\,. \tag{118}$$

If $x \notin \mathcal{N}_{\mathcal{C},t}$, then every one-coordinate neighbor of x has distance at least R_t from $\mathcal{C}$. Summing Eq. (118) over $x \notin \mathcal{N}_{\mathcal{C},t}$ and $i \in [d]$ gives

$$\varepsilon_{\mathsf{unif}}(\mathcal{O}; q^{\mathcal{C}}, t) \leqslant d\left(1 + \log\frac{1-\beta_t}{\beta_t}\right)\mathbb{P}[\mathrm{Bin}(d,\beta_t) \geq R_t]\,. \tag{119}$$

If $\beta_t \geq d^{-2}$, then the logarithmic factor in Eq. (119) is at most $C(1+\log d)$, and Hoeffding's inequality proves Eq. (115). If $0 < \beta_t < d^{-2}$, set $m = \lceil \gamma dw \rceil$. By the Chernoff bound,

$$\mathbb{P}[\mathrm{Bin}(d,\beta_t) \geq R_t] \leqslant \left(\frac{ed\beta_t}{m}\right)^m\,.$$

For $m \geq 2$, the function $u \mapsto (1+\log(1/u))u^m$ is increasing on $(0, d^{-2}]$. Therefore,

$$\left(1 + \log\frac{1-\beta_t}{\beta_t}\right)\mathbb{P}[\mathrm{Bin}(d,\beta_t) \geq R_t] \leqslant C(1+\log d)\left(\frac{e}{dm}\right)^m \leqslant C(1+\log d)e^{-cdw^2}\,. \tag{120}$$

The last inequality uses $m \geq \gamma dw$, $w \leqslant 1$, and hence $m\log(dm/e) \geq cdw\log d \geq cdw^2$ for sufficiently large d. Eqs. (119)–(120) prove Eq. (115) uniformly over t.

Now fix x before drawing $\mathcal{C}$. Since $\mathcal{C}$ is a uniform 2^k-element subset of $\mathcal{X}$,

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathcal{C}}[x \in \mathcal{N}_{\mathcal{C},t}] &\leqslant \mathbb{E}_{\mathcal{C}}\,|\mathcal{C} \cap \{y : d_H(x,y) \leqslant R_t\}| \\ &= \frac{2^k}{2^d}\sum_{j=0}^{R_t}\binom{d}{j} \leqslant \exp\left(k\log 2 - d\log 2 + dh_2(R_t/d)\right)\,.\end{aligned} \tag{121}$$

By Eq. (106),

$$k\log 2 = d\,\mathsf{I}_{\mathsf{unif}}(t^\star_{\mathsf{unif}}) = d\big(\log 2 - h_2(\beta_{t^\star_{\mathsf{unif}}})\big)\,. \tag{122}$$

Moreover, $h_2'(u) = \log((1-u)/u)$ is bounded below by a positive constant for $u \leqslant \beta_{\tau_+} < 1/2$. Eq. (117) consequently gives

$$h_2(\beta_{t^\star_{\mathsf{unif}}}) - h_2(R_t/d) = \int_{R_t/d}^{\beta_{t^\star_{\mathsf{unif}}}} \log\frac{1-u}{u}\,\mathrm{d}u \geq cw\,. \tag{123}$$

Substituting Eqs. (122) and (123) into Eq. (121) proves Eq. (116). □

The oracle now follows the three regimes suggested by the two lemmas above. It gives the exact score within distance $4w$ of $t^\star_{\mathsf{unif}}$ (the window). Below this window it gives the exact score only on $\mathcal{N}_{\mathcal{C},t}$ and all-one score otherwise. Above the window it gives the all-one score whenever that answer has error at most η, and otherwise gives the exact score.

Definition I.4 (Adversarial uniform diffusion oracle)**.** *For each fixed instance $(k, \mathcal{C})$, define $\mathcal{O}_{k,\mathcal{C}}$ by*

$$\mathcal{O}_{k,\mathcal{C}}(t,x) := \begin{cases} s_t^{\mathcal{C}}(x), & |t - t^\star_{\mathsf{unif}}| < 4w, \\ s_t^{\mathcal{C}}(x), & t \leqslant t^\star_{\mathsf{unif}} - 4w \text{ and } x \in \mathcal{N}_{\mathcal{C},t}, \\ \mathbf{1}, & t \leqslant t^\star_{\mathsf{unif}} - 4w \text{ and } x \notin \mathcal{N}_{\mathcal{C},t}, \\ \mathbf{1}, & t \geq t^\star_{\mathsf{unif}} + 4w \text{ and } \varepsilon_{\mathsf{unif}}(\mathcal{O}_0; q^{\mathcal{C}}, t) \leqslant \eta, \\ s_t^{\mathcal{C}}(x), & t \geq t^\star_{\mathsf{unif}} + 4w \text{ and } \varepsilon_{\mathsf{unif}}(\mathcal{O}_0; q^{\mathcal{C}}, t) > \eta. \end{cases}.$$

Proposition I.5 (Validity and indistinguishability of the oracle)**.** *The constants in Theorem I.1 can be chosen so that the oracles in Definition I.4 satisfy Eq. (110) for every allowed instance. For these choices, whenever $|t - t^\star_{\mathsf{unif}}| \geq 4w$ and $x \in \mathcal{X}$ is fixed independently of $\mathcal{C}$,*

$$\mathop{\mathbb{P}}_{\mathcal{C}}[\mathcal{O}_{k,\mathcal{C}}(t,x) \neq \mathcal{O}_0(t,x)] \leqslant e^{-4\Lambda}. \tag{124}$$

Proof. We first choose A large enough that Lemma I.2 holds with $K = 6$ and that, by Lemma I.3 (whose hypothesis $\log d \leqslant dw^2 = A^2\Lambda \leqslant c_1 d$ holds once $c_0 \leqslant c_1/A^2$), uniformly over $t \leqslant t^\star_{\mathsf{unif}}(k) - 4w$,

$$\varepsilon_{\mathsf{unif}}(\mathcal{O}_{k,\mathcal{C}}; q^{\mathcal{C}}, t) \leqslant Cd(1+\log d)e^{-cdw^2} = Cd(1+\log d)e^{-cA^2\Lambda} \leqslant \eta, \tag{125}$$

once A and then d_0 are sufficiently large. We then further decrease c_0 and increase d_0 so that, by Lemma I.3 again, uniformly over $t \leqslant t^\star_{\mathsf{unif}}(k) - 4w$,

$$\mathbb{E}_{\mathcal{C}}\, \mathbb{1}[x \in \mathcal{N}_{\mathcal{C},t}] \leqslant e^{-cdw} \leqslant e^{-4\Lambda}. \tag{126}$$

Fix an instance and a time. In the interval $|t - t^\star_{\mathsf{unif}}| < 4w$, Definition I.4 returns the exact score. If $t \leqslant t^\star_{\mathsf{unif}} - 4w$, Eq. (125) gives error at most η. If $t \geq t^\star_{\mathsf{unif}} + 4w$, the oracle returns the all-one score only when its error is at most η and returns the exact score otherwise. This proves Eq. (110).

It remains to prove Eq. (124). Below the window, the oracle can differ from $\mathcal{O}_0$ only when $x \in \mathcal{N}_{\mathcal{C},t}$, so Eq. (126) gives the desired bound. Above the window, the two answers can differ only when $\varepsilon_{\mathsf{unif}}(\mathcal{O}_0; q^{\mathcal{C}}, t) > \eta$. Markov's inequality and Lemma I.2 with $K = 6$ give

$$\mathop{\mathbb{P}}_{\mathcal{C}}[\varepsilon_{\mathsf{unif}}(\mathcal{O}_0; q^{\mathcal{C}}, t) > \eta] \leqslant \frac{e^{-6\Lambda}}{\eta} \leqslant e^{-4\Lambda},$$

where the last step uses $\eta \geq e^{-2\Lambda} = (\eta\rho^2/(C_\Lambda d))^2$, which holds since $\eta, \rho < 1$ and we may take $C_\Lambda \geq 1$. This finishes the proof of the proposition. □

I.2 Distribution indistinguishability: the lower bound proof

Let R denote all the internal randomness of an adaptive algorithm $\mathcal{A}$. The following compares the algorithm's interaction with the constructed oracle to its interaction with $\mathcal{O}_0$. For the latter, every answer is the all-one score, so the entire query sequence is determined by R and is independent of k and $\mathcal{C}$.

Proposition I.6 (Coupling argument)**.** *Couple an algorithm's run against $\mathcal{O}_{k,\mathcal{C}}$ with its run against $\mathcal{O}_0$, using the same internal randomness R. For an algorithm making at most Q queries, let $p_{\mathrm{diff}}(k,\mathcal{C})$ be the conditional probability over R given $k, \mathcal{C}$ that the runs receive different oracle answers at some point. Then*

$$\mathbb{E}_{k,\mathcal{C}} p_{\mathrm{diff}}(k,\mathcal{C}) \leqslant Q\left(\frac{1}{|\mathcal{K}_d^{\mathsf{U}}|} + e^{-4\Lambda}\right). \tag{127}$$

Proof. Let $\mathcal{D}$ denote the event that the two coupled runs receive different oracle answers at some point. By definition,

$$\mathbb{E}_{k,\mathcal{C}} p_{\mathrm{diff}}(k,\mathcal{C}) = \mathbb{E}_{k,\mathcal{C}} \mathop{\mathbb{P}}_{R}[\mathcal{D} \mid k, \mathcal{C}] = \mathop{\mathbb{P}}_{k,\mathcal{C},R}[\mathcal{D}] = \mathbb{E}_R \mathop{\mathbb{P}}_{k,\mathcal{C}}[\mathcal{D} \mid R]. \tag{128}$$

Now condition on any fixed realization of the internal randomness $R = r$. We will bound $\mathbb{P}_{k,\mathcal{C}}[\mathcal{D} \mid R = r]$ uniformly over r.

It is enough to bound the probability of a first different answer, because the queries agree up to that point.

Run the algorithm against $\mathcal{O}_0$, padding with irrelevant queries if it stops before making Q queries. This produces a deterministic sequence $(t_j, x_j)_{j=1}^{Q}$, independent of k and $\mathcal{C}$. Consider the randomness of k. By Eq. (108), the possible values of $t^\star_{\mathsf{unif}}(k)$ are separated by at least $10w$. Therefore, for each query time t_j, the interval $|t_j - t^\star_{\mathsf{unif}}(k)| < 4w$ contains the critical time of at most one $k \in \mathcal{K}_d^{\mathsf{U}}$, and hence

$$\mathop{\mathbb{P}}_{k}[|t_j - t^\star_{\mathsf{unif}}(k)| < 4w] \leqslant \frac{1}{|\mathcal{K}_d^{\mathsf{U}}|}\,.$$

A union bound shows that at least one query time lies in the critical window with probability at most $Q/|\mathcal{K}_d^{\mathsf{U}}|$ over k.

Fix a k for which every query time lies outside the critical window, i.e., $|t_j - t^\star_{\mathsf{unif}}(k)| \geq 4w$ for every j. Since each x_j was fixed before $\mathcal{C}$ was drawn, Proposition I.5 gives, for every j,

$$\mathop{\mathbb{P}}_{\mathcal{C}}[\mathcal{O}_{k,\mathcal{C}}(t_j, x_j) \neq \mathcal{O}_0(t_j, x_j)] \leqslant e^{-4\Lambda}\,.$$

A union bound over the Q queries shows that some oracle answer differs outside the window with probability at most $Qe^{-4\Lambda}$.

Consequently, for the fixed realization $R = r$, the probability that a query enters the critical window or that an oracle answer differs outside the window is at most the right-hand side of Eq. (127). If neither event occurs, the two runs have identical queries and answers and hence make the same next query at every step, so $\mathcal{D}$ does not occur. Averaging over R in Eq. (128) proves the proposition. □

Proof of Theorem I.1. Proposition I.5 gives an oracle satisfying Eq. (110) for every instance. Fix an adaptive randomized algorithm $\mathcal{A}$, and let ν be its output law when every answer comes from $\mathcal{O}_0$. Since $\mathcal{O}_0$ gives the same answer on every instance, ν is independent of k and $\mathcal{C}$.

For each fixed instance $(k, \mathcal{C})$, couple the run against $\mathcal{O}_{k,\mathcal{C}}$ with the run against $\mathcal{O}_0$ using the same internal randomness R. Let $Y_{k,\mathcal{C}}(R)$ and $Y_0(R)$ denote their outputs. Their laws over R are $\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}$ and ν, respectively, and hence

$$\begin{aligned}\mathsf{TV}(\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}, \nu) &\leqslant \mathop{\mathbb{P}}_{R}[Y_{k,\mathcal{C}}(R) \neq Y_0(R)] \\ &\leqslant p_{\mathrm{diff}}(k, \mathcal{C})\,.\end{aligned}$$

Since $q^{\mathcal{C}}$ is supported on $\mathcal{C}$,

$$\mathsf{TV}(\widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}, q^{\mathcal{C}}) \geq 1 - \widehat{q}_{\mathcal{A}}^{\mathcal{O}_{k,\mathcal{C}}}(\mathcal{C}) \geq 1 - \nu(\mathcal{C}) - p_{\mathrm{diff}}(k, \mathcal{C})\,. \tag{129}$$

Fix k. A fixed point of $\mathcal{X}$ belongs to a uniformly chosen 2^k-element subset with probability 2^{k-d}. Since ν is independent of $\mathcal{C}$,

$$\mathbb{E}_{\mathcal{C}|k}\nu(\mathcal{C}) = \mathbb{E}_{X\sim\nu}\mathop{\mathbb{P}}_{\mathcal{C}|k}[X \in \mathcal{C}] = 2^{k-d} \leqslant e^{-(\log 2 - \kappa_+)d}\,. \tag{130}$$

We now choose the constant c in Eq. (111). By Eqs. (109) and (111),

$$\frac{Q}{|\mathcal{K}_d^{\mathsf{U}}|} \leqslant \frac{cA}{c_{\mathsf{U}}}\rho^2\,.$$

Choose $c \leqslant c_{\mathsf{U}}/(4A)$, so this term is at most $\rho^2/4$. Since $\Lambda \geq \log d$,

$$Qe^{-4\Lambda} \leqslant c\rho^2\sqrt{\frac{d}{\Lambda}}\,e^{-4\Lambda} \leqslant c\rho^2 e^{-7\Lambda/2} \leqslant \frac{\rho^2}{4}$$

after increasing d_0. Increase d_0 once more so that the right-hand side of Eq. (130) is at most $\rho^2/2$. Proposition I.6 then gives

$$\mathbb{E}_{k,\mathcal{C}}\big[\nu(\mathcal{C}) + p_{\mathrm{diff}}(k, \mathcal{C})\big] \leqslant \rho^2\,.$$

Markov's inequality yields

$$\mathop{\mathbb{P}}_{k,\mathcal{C}}[\nu(\mathcal{C}) + p_{\mathrm{diff}}(k, \mathcal{C}) \leqslant \rho] \geq 1 - \rho\,.$$

Eq. (129) proves Eq. (112). Finally, $\Lambda = O(\log d)$ for polynomially small η and constant ρ, giving the stated $\Omega(\sqrt{d/\log d})$ lower bound. □

J Deferred proofs

J.1 Proofs from Section 2

J.1.1 Proof of Lemma B.1

Proof. Fix $y \in \mathcal{X}$ and $i \in [d]$, and denote

$$v \coloneqq s_t(y)[i, \cdot]\,, \qquad V \coloneqq \sum_{b \in \Sigma} v_b\,.$$

Note that $v_{y_i} = 1$. Let

$$g_a \coloneqq \mathbb{P}[X_s^i = a,\ (X_t)_{-i} = y_{-i}]\,, \qquad G \coloneqq \sum_{a \in \Sigma} g_a = \mathbb{P}[(X_t)_{-i} = y_{-i}]\,.$$

Conditionally on X_s, the coordinates of X_t are independent, and $(X_t)_i \sim K_h^{\mathsf{U}}(\cdot \mid X_s^i)$. As a result, for every $b \in \Sigma$,

$$q_t(y^{i \leftarrow b}) = \sum_{a \in \Sigma} g_a K_h^{\mathsf{U}}(b \mid a) = \alpha_h g_b + \beta_h G\,. \tag{131}$$

Summing Eq. (131) over $b \in \Sigma$ and using $\alpha_h + S\beta_h = 1$ gives $\sum_{b \in \Sigma} q_t(y^{i \leftarrow b}) = G$, so dividing by $q_t(y)$ identifies the normalizing constant as

$$V = G / q_t(y)\,.$$

The Gibbs sampling marginal. Dividing Eq. (131) by $q_t(y)$ and then by $V = G/q_t(y)$,

$$\frac{v_b}{V} = \alpha_h \cdot \frac{g_b}{G} + \beta_h\,. \tag{132}$$

As $g_b/G = \mathbb{P}[X_s^i = b \mid (X_t)_{-i} = y_{-i}]$, solving Eq. (132) for g_b/G gives the second claim.

The posterior marginal. Write $\pi \coloneqq p_h^{\mathsf{gibbs}}[v]$ for the distribution just computed. As $(X_t)_i$ is conditionally independent of $(X_t)_{-i}$ given X_s^i, by Bayes' rule we have

$$\mathbb{P}[X_s^i = a \mid X_t = y] = \frac{\pi_a K_h^{\mathsf{U}}(y_i \mid a)}{\sum_{a' \in \Sigma} \pi_{a'} K_h^{\mathsf{U}}(y_i \mid a')}\,.$$

The denominator equals $\alpha_h \pi_{y_i} + \beta_h$, which by Eq. (132) is $v_{y_i}/V = 1/V$. Substituting this and $\alpha_h V \pi_a = v_a - \beta_h V$, both read off from Eq. (132), we obtain the first claim,

$$\mathbb{P}[X_s^i = a \mid X_t = y] = V\pi_a\big(\alpha_h\, \mathbb{1}[a = y_i] + \beta_h\big) = \frac{\alpha_h\, \mathbb{1}[a = y_i] + \beta_h}{\alpha_h}\big(v_a - \beta_h V\big) = p_{h, y_i}^{\mathsf{post}}[v](a)\,.$$

□

J.1.2 Proof of Lemma B.6

Proof. Fix I with $|I| = m$, X_0, and $i \notin I$, and write $Y = X_0^{(I)}$; the partial assignment $(I, X_{0,I})$ is consistent. By Lemma B.4, the definition of $\widehat{q}^{\mathcal{O}}_{i|I}$, and the homogeneity $\psi(\lambda c, \lambda s) = \lambda \psi(c, s)$ for $\lambda > 0$,

$$\begin{aligned}
&\frac{1 - \alpha_t}{\alpha_t} \sum_{a \in [S]} \psi(s_t(Y)[i, a], \widehat{s}_t(Y)[i, a]) \\
&= \sum_{a \in [S]} \left(\widehat{q}^{\mathcal{O}}_{i|I}(a \mid X_{0,I}) - q_{i|I}(a \mid X_{0,I}) + q_{i|I}(a \mid X_{0,I}) \log \frac{q_{i|I}(a \mid X_{0,I})}{\widehat{q}^{\mathcal{O}}_{i|I}(a \mid X_{0,I})} \right) \\
&= \mathsf{KL}(q_{i|I}(\cdot \mid X_{0,I}) \,\|\, \widehat{q}^{\mathcal{O}}_{i|I}(\cdot \mid X_{0,I}))\,,
\end{aligned}$$

since $\sum_{a \in [S]} q_{i|I}(a \mid X_{0,I}) = \sum_{a \in [S]} \widehat{q}^{\mathcal{O}}_{i|I}(a \mid X_{0,I}) = 1$. Averaging over $i \notin I$ and taking the expectation over I and X_0 gives Eq. (2). □

J.2 Proofs from Section C

J.2.1 Proof of Proposition C.6

Proof. For the first part, write $v_a \coloneqq s_t(y)[i,a] = q_t(y^{i\leftarrow a})/q_t(y)$. Every entry of $K_t^{\mathsf{U}}(\cdot \mid z)$ is at least $\beta_t > 0$, so q_t has full support and $v \in \mathbb{R}^S_{>0}$; moreover $v_{y_i} = 1$ since $y^{i\leftarrow y_i} = y$. For the constraint $v_a / \sum_{a'\in\Sigma} v_{a'} \geq \beta_t$,

$$\frac{v_a}{\sum_{a'\in\Sigma} v_{a'}} = \frac{q_t(y^{i\leftarrow a})}{\sum_{a'\in\Sigma} q_t(y^{i\leftarrow a'})} = \mathbb{P}[(X_t)_i = a \mid (X_t)_{-i} = y_{-i}]\,.$$

That this is lower bounded by β_t follows by Proposition C.3.

For the second part, write $w_a \coloneqq v_a / \sum_{a'\in\Sigma} v_{a'}$, so that $p \coloneqq p_h^{\mathsf{gibbs}}[v]$ is given by $p(a) = (w_a - \beta_h)/\alpha_h$. It is immediate that $\sum_a p(a) = 1$. Moreover, $w_a \geq \beta_t$ by definition, and $\beta_t - \beta_h = \frac{e^{-h}-e^{-t}}{S} = \alpha_h\beta_s$, so $p(a) \geq (\beta_t - \beta_h)/\alpha_h = \beta_s$. Finally, the identity $p_{h,b}^{\mathsf{post}}[v](a) = \big(\sum_{a'} v_{a'}\big) K_h^{\mathsf{U}}(b \mid a)\, p(a)$ follows from the definitions, and summing it over a gives $\big(\sum_{a'} v_{a'}\big)(\alpha_h p(b) + \beta_h) = v_b = 1$. □

J.2.2 Proof of Lemma C.10

Proof. For each coordinate $i \in [d]$, let $B_i \sim \mathrm{Ber}(e^{-T})$ be independent, and let $U = (U_1, \ldots, U_d) \sim \mathrm{Unif}(\mathcal{X})$ be independent of $X_0 \sim q$ and B. Define the coupled random vector $\widetilde{X}$ by

$$\widetilde{X}_i \coloneqq \begin{cases} (X_0)_i, & B_i = 1, \\ U_i, & B_i = 0. \end{cases}.$$

Then $\widetilde{X} \sim q_T$ (this is exactly the coordinatewise kernel $(K_T^{\mathsf{U}})^{\otimes d}$ with $\alpha_T = e^{-T}$ and $\beta_T = (1 - e^{-T})/S$).

Conditioned on $B = b \in \{0,1\}^d$, the coordinates with $b_i = 0$ are i.i.d. uniform, and the coordinates with $b_i = 1$ have joint law equal to the corresponding marginal of q. Let $q^{(b)}$ denote the conditional law of $\widetilde{X}$ given $B = b$. By convexity of KL in its first argument,

$$\mathsf{KL}(q_T \parallel \mathrm{Unif}(\mathcal{X})) = \mathsf{KL}(\mathbb{E}_B[q^{(B)}] \parallel \mathrm{Unif}(\mathcal{X})) \leqslant \mathbb{E}_B\, \mathsf{KL}(q^{(B)} \parallel \mathrm{Unif}(\mathcal{X}))\,.$$

For any fixed b, write $I(b) \coloneqq \{i : b_i = 1\}$ and $k \coloneqq |I(b)|$. Since $q^{(b)}$ is uniform on coordinates outside $I(b)$,

$$\mathsf{KL}(q^{(b)} \parallel \mathrm{Unif}(\mathcal{X})) = \mathsf{KL}(\mathrm{law}(X_{0,I(b)}) \parallel \mathrm{Unif}(\Sigma^{I(b)})) \leqslant k \log S\,.$$

Taking expectation over B gives

$$\mathsf{KL}(q_T \parallel \mathrm{Unif}(\mathcal{X})) \leqslant \mathbb{E}_B|I(B)| \log S = de^{-T} \log S\,,$$

as claimed. □

J.2.3 Proof of Lemma C.11

Proof. Fix $0 < t \leqslant 1$ and use the same mask coupling as in the proof of Lemma C.10, with $B_i \sim \mathrm{Ber}(e^{-t})$ and $\widetilde{X} \sim q_t$ defined by

$$\widetilde{X}_i = \begin{cases} (X_0)_i, & B_i = 1, \\ U_i, & B_i = 0, \end{cases},$$

where $U \sim \mathrm{Unif}(\Sigma^d)$ is independent.

Using $H(\widetilde{X}) \leqslant H(\widetilde{X}, B) = H(B) + H(\widetilde{X} \mid B)$, we obtain

$$H(q_t) - H(q) = H(\widetilde{X}) - H(X_0) \leqslant H(B) + \big(H(\widetilde{X} \mid B) - H(X_0)\big)\,.$$

Conditioned on $B = b$, the refreshed coordinates contribute $(d - |I(b)|) \log S$ nats of entropy and are independent of everything else, while the kept coordinates are a function of X_0. Hence

$$H(\widetilde{X} \mid B = b) = H(X_{0,I(b)}) + (d - |I(b)|) \log S \leqslant H(X_0) + (d - |I(b)|) \log S \,.$$

Averaging over B yields

$$H(\widetilde{X} \mid B) - H(X_0) \leqslant \mathbb{E}_B[d - |I(B)|] \log S = d(1 - e^{-t}) \log S \leqslant dt \log S \,.$$

It remains to bound $H(B)$. Since B has i.i.d. coordinates with $\mathbb{P}[B_i = 0] = 1 - e^{-t}$, we have

$$H(B) = d\, h_2(1 - e^{-t}) \,.$$

For $0 < \rho \leqslant 1$, the standard bound $h_2(\rho) \leqslant \rho \log(e/\rho)$ gives

$$H(B) \leqslant d(1 - e^{-t}) \log\Big(\frac{e}{1 - e^{-t}}\Big) \,.$$

For $0 < t \leqslant 1$ we have $1 - e^{-t} \leqslant t$, and also $\rho \mapsto \rho \log(e/\rho)$ is increasing on $(0, 1]$, so

$$H(B) \leqslant dt \log(e/t) \,.$$

Combining the two parts,

$$H(q_t) - H(q) \leqslant dt \log(e/t) + dt \log S = dt \log(eS/t) \,,$$

as desired. □

J.3 Proofs from Section D

J.3.1 Proof of Lemma D.6

Proof. Apply Lemma D.5 with $\mu_u = f_{\alpha,u}$, $\nu_u = f_{\beta,u}$, and $\ell_u = \log(f_{\alpha,u}/f_{\beta,u})$; the first part is then the first identity of that lemma. For the second part, since $\nabla \log f_{\lambda,u}(z) = (\mathsf{m}_{\lambda,u}(z) - z)/u$ for every $\lambda \in \Delta(\Sigma)$, a direct calculation yields

$$\nabla \ell_u = \nabla \log f_{\alpha,u} - \nabla \log f_{\beta,u} = \frac{1}{u}(\mathsf{m}_{\alpha,u}(z) - \mathsf{m}_{\beta,u}(z)) \,,$$

$$\nabla^2 \log \nu_u = \nabla^2 \log f_{\beta,u} = -\frac{1}{u}\mathrm{Id}_S + \frac{1}{u^2}\mathsf{C}_{\beta,u}(z) \,,$$

$$\nabla^2 \ell_u = \nabla^2 \log f_{\alpha,u} - \nabla^2 \log f_{\beta,u} = \frac{1}{u^2}(\mathsf{C}_{\alpha,u}(z) - \mathsf{C}_{\beta,u}(z)) \,.$$

By Lemma D.5 and the fact that $\mathsf{C}_{\lambda,u}(z) \succeq 0$ for all λ,

$$\begin{aligned}\partial_u \mathsf{FI}(f_{\alpha,u} \parallel f_{\beta,u}) &\geq -\frac{1}{u^4}\mathbb{E}_{f_{\alpha,u}}\|\mathsf{C}_{\alpha,u}(z) - \mathsf{C}_{\beta,u}(z)\|_F^2 - \frac{2}{u^3}\mathbb{E}_{f_{\alpha,u}}\|\mathsf{m}_{\alpha,u}(z) - \mathsf{m}_{\beta,u}(z)\|_2^2 \\ &\geq -\Big(\frac{2}{u} + \frac{3}{u^2}\Big)\mathsf{FI}(f_{\alpha,u} \parallel f_{\beta,u}) \,,\end{aligned}$$

where in the last step we used Proposition J.1 below. □

The above proof used the following elementary fact:

Proposition J.1. *Given $p \in \Delta(\Sigma)$, define $\mathsf{C}(p) = \mathrm{diag}(p) - p^{\otimes 2}$. Then for all $p, p' \in \Delta(\Sigma)$,*

$$\|\mathsf{C}(p) - \mathsf{C}(p')\|_F^2 \leqslant 3\|p - p'\|_2^2 \,.$$

Proof. Let $s \coloneqq p + p'$ and $\delta \coloneqq p - p'$. Then $\mathsf{C}(p) - \mathsf{C}(p') = \mathrm{diag}(\delta) - \frac{1}{2}(s\delta^\intercal + \delta s^\intercal)$, so

$$\|\mathsf{C}(p) - \mathsf{C}(p')\|_F^2 = \|\delta\|_2^2 - 2\sum_a s_a \delta_a^2 + \frac{1}{2}\|s\|_2^2\|\delta\|_2^2 + \frac{1}{2}\langle s, \delta\rangle^2 \leqslant 3\|\delta\|_2^2 \,,$$

as $\|s\|_2^2 \leqslant 4$ and $\langle s, \delta\rangle^2 \leqslant 2\sum_a s_a \delta_a^2$ by Cauchy–Schwarz. □

J.3.2 Proof of Lemma D.7

Proof. Write φ_a for the density of $N(e_a, \rho \mathrm{Id}_S)$ and $h \coloneqq \alpha - \beta$. We have

$$\mathsf{KL}(f_{\alpha,\rho} \parallel f_{\beta,\rho}) \leqslant \chi^2\left(f_{\alpha,\rho} \| f_{\beta,\rho}\right) \leq \int \frac{(\sum_a h_a \varphi_a)^2}{\sum_a \beta_a \varphi_a}\,.$$

For fixed h, the right-hand side is convex in β, so it is maximized at some vertex $\beta = e_b$, in which case the denominator is φ_b. A direct calculation gives

$$\int \frac{\varphi_a(z)\varphi_c(z)}{\varphi_b(z)}\,\mathrm{d}z = \exp\left(\frac{(e_a - e_b)^\intercal (e_c - e_b)}{\rho}\right)\,.$$

Using $\sum_a h_a = 0$ and expanding the quadratic form, with $x = e^{1/\rho}$, we conclude that

$$\int \frac{(\sum_a h_a \varphi_a)^2}{\varphi_b} = (x-1)h_b^2 + (x^2 - x)\sum_{a \neq b} h_a^2 \leq (x^2-1)\|h\|_2^2\,,$$

as claimed. □

J.4 Proofs from Section F

Throughout this subsection we use the notation of Section F: Λ and $w = A\sqrt{\Lambda/d}$ are as in Eqs. (24) and (26), $I_0 = [\tau_-/2, \tau_+ + 1]$ is the interval of Eq. (23), $L_t = q_t^{\mathcal{C}}/2^{-d}$ is the likelihood ratio with respect to the uniform distribution, $\ell_t = \beta_t/(1-\beta_t)$, $\widetilde{s}_t$ is the clipped approximate score of Eq. (33), $\widehat{T}_t$ is the statistic of Eq. (34), and F_t is the squared clipped score error of Eq. (35).

J.4.1 Proof of Lemma F.2

Proof. Define the truncated likelihood ratios

$$\widetilde{L}_z^c(x) = 2^d (K_t^{\mathsf{U}})^{\otimes d}(x \mid z) \cdot \mathbb{1}[d_H(x,z) > r] \quad \text{and} \quad \widetilde{L}_z(x) = 2^d (K_t^{\mathsf{U}})^{\otimes d}(x \mid z) \cdot \mathbb{1}[d_H(x,z) \leqslant r]\,.$$

We can write the likelihood ratio as

$$L_t(x) = \frac{q_t^{\mathcal{C}}(x)}{2^{-d}} = \mathbb{E}_{z \sim \mathrm{Unif}(\mathcal{C})}\, \widetilde{L}_z^c(x) + \mathbb{E}_{z \sim \mathrm{Unif}(\mathcal{C})}\, \widetilde{L}_z(x)\,.$$

Also define $\mu(x) = \mathbb{E}_{z \sim \mathrm{Unif}(\mathcal{X})}\, \widetilde{L}_z^c(x)$. Note that $\mu(x) \leqslant 1$ because the expectation of the (untruncated) likelihood ratio is 1. Then

$$2\mathbb{E}_{\mathcal{C}} \mathsf{TV}(q_t^{\mathcal{C}}, \mathrm{Unif}(\mathcal{X})) \leqslant \underbrace{\mathbb{E}_{\mathcal{C}} \mathbb{E}_{x \sim \mathrm{Unif}(\mathcal{X})} \left| \mathbb{E}_{z \sim \mathrm{Unif}(\mathcal{C})}\, \widetilde{L}_z^c(x) - \mu(x) \right|}_{①} + \underbrace{\mathbb{E}_{x \sim \mathrm{Unif}(\mathcal{X})}[1 - \mu(x)]}_{②}$$
$$+ \underbrace{\mathbb{E}_{\mathcal{C}} \mathbb{E}_{x \sim \mathrm{Unif}(\mathcal{X})} \mathbb{E}_{z \sim \mathrm{Unif}(\mathcal{C})}\, \widetilde{L}_z(x)}_{③}\,.$$

We have

$$\begin{aligned}
①^2 &\leqslant \mathbb{E}_{\mathcal{C}} \mathbb{E}_{x \sim \mathrm{Unif}(\mathcal{X})} \left(\mathbb{E}_{z \sim \mathrm{Unif}(\mathcal{C})}\, \widetilde{L}_z^c(x) - \mu(x) \right)^2 \\
&\leqslant \frac{1}{M} \mathbb{E}_{x,z \sim \mathrm{Unif}(\mathcal{X})} \left(\widetilde{L}_z^c(x) - \mu(x) \right)^2 \\
&\leqslant \frac{1}{M} \mathbb{E}_{x,z \sim \mathrm{Unif}(\mathcal{X})}\, \widetilde{L}_z^c(x)^2 \\
&\leqslant \frac{1}{M} (1 + e^{-t})^d \cdot \tanh(t/2)^{(r+1)}\,,
\end{aligned}$$

where in the second step we used that the variance of the empirical average over samples drawn without replacement is at most the variance of the empirical average over samples drawn with replacement, and in the last step we used that $\widetilde{L}_z^c(x)^2 \leqslant (1 + e^{-t})^d \tanh(t/2)^{r+1} \cdot 2^d (K_t^{\mathsf{U}})^{\otimes d}(x \mid z)$ by Lemma B.2, together with $\mathbb{E}_{x,z \sim \mathrm{Unif}(\mathcal{X})} 2^d (K_t^{\mathsf{U}})^{\otimes d}(x \mid z) = 1$. Finally,

$$② = ③ = \mathbb{E}_{x,z \sim \mathrm{Unif}(\mathcal{X})} \widetilde{L}_z(x) = \mathop{\mathbb{P}}_{z \sim \mathrm{Unif}(\mathcal{X}), x \sim (K_t^{\mathsf{U}})^{\otimes d}(\cdot \mid z)} [d_H(x,z) \leqslant r]\,.$$

The claimed bound follows upon observing that the above quantity is $\mathbb{P}[\mathrm{Bin}(d, \beta_t) \leqslant r]$. □

J.4.2 Proof of Lemma F.3

Proof. We will bound the Bayes-optimal recovery error by the error of the nearest-neighbor decoder, i.e., the one which given x outputs anything in $\arg\min_{z' \in \mathcal{C}} d_H(x, z')$. This decoder succeeds when the true z satisfies $d_H(x, z) \leqslant r$ and for all $z' \in \mathcal{C} \backslash \{z\}$, $d_H(x, z') > r$.

The probability that $d_H(x, z) > r$ is exactly the second term on the right-hand side of the claimed bound, because the distribution of z given by sampling random $\mathcal{C}$ and then taking $z \sim q^{\mathcal{C}}$ is uniform over $\mathcal{X}$, and conditioned on z the quantity $d_H(x, z)$ is distributed as $\mathrm{Bin}(d, \beta_t)$.

The probability that some $z' \in \mathcal{C} \setminus \{z\}$ has $d_H(x, z') \leqslant r$ is at most the expected number of such z'. Conditioned on any x, z, the remaining elements of $\mathcal{C}$ are a random $(M-1)$-sized subset of the cube, so the expected number of such z' is at most $\frac{M-1}{2^d-1} \cdot |B_r(x)|$, where $B_r(x)$ is the closed Hamming ball of radius r around x. The size of $B_r(x)$ is independent of x and equal to $2^d \, \mathbb{P}[\mathrm{Bin}(d, 1/2) \leqslant r]$, so the claim follows. □

J.4.3 Proof of Lemma F.6

Proof. For the first part of Eq. (36), the definitions of T_t and $\widehat{T}_t$ (Eq. (34)) and the Cauchy–Schwarz inequality give

$$|\widehat{T}_t(x) - T_t(x)| \leqslant \frac{1 - e^{-2t}}{2e^{-t}} \sum_{i=1}^{d} \left| \widetilde{s}_t(x)[i, -x_i] - s_t(x)[i, -x_i] \right| \leqslant \frac{1 - e^{-2t}}{2e^{-t}} \sqrt{d F_t(x)} \,,$$

and the prefactor $\frac{1-e^{-2t}}{2e^{-t}}$ is bounded on I_0 by a constant depending only on $[\kappa_-, \kappa_+]$.

For the second part, fix any entry $c := s_t(x)[i, -x_i]$ of the true score. The function $u \mapsto \psi(c, u)$ decreases on $(0, c]$ and increases on $[c, \infty)$, since its derivative is $1 - c/u$. By Eq. (32), $c \in [\ell_t, \ell_t^{-1}]$, so clipping the estimated entry $\widehat{s}_t(x)[i, -x_i]$ to this interval cannot increase $\psi(c, \cdot)$. On the compact time interval I_0, all the clipping intervals $[\ell_t, \ell_t^{-1}]$ lie in a single interval $[m_0, M_0] \subset (0, \infty)$. For $c, u \in [m_0, M_0]$, the second derivative of $u \mapsto \psi(c, u)$, which is c/u^2, is at least m_0/M_0^2. Taylor's theorem at $u = c$ gives

$$(u - c)^2 \leqslant \frac{2M_0^2}{m_0} \psi(c, u) \,.$$

Applying this inequality to each clipped entry $u = \widetilde{s}_t(x)[i, -x_i]$, summing over $i \in [d]$, and averaging under $q_t^{\mathcal{C}}$ gives the second part of Eq. (36) with $C' = 2M_0^2/m_0$, which depends only on I_0 and hence only on $[\kappa_-, \kappa_+]$. □

J.4.4 Proof of Lemma F.7

Proof. Consider a binomial random variable with d trials and success probability $v = r/d$. Its most likely value is r, so the probability at r lies between $1/(d+1)$ and 1. Since this probability is $\binom{d}{r} e^{-d h_2(v)}$, Eq. (37) follows.

For a different success probability p, multiplying the lower bound in Eq. (37) by $p^r (1-p)^{d-r}$ gives the first inequality in Eq. (38). The second follows from Eq. (39), which holds for all $v, p \in (0, 1)$ by applying $\log u \leqslant u - 1$ to the two terms of the KL divergence:

$$\mathsf{KL}(\mathrm{Ber}(v) \parallel \mathrm{Ber}(p)) \leqslant \frac{v^2}{p} + \frac{(1-v)^2}{1-p} - 1 = \frac{(v-p)^2}{p(1-p)} \,. \qquad \square$$

J.4.5 Proof of Lemma F.8

Proof. Choose $r = \lceil d \max\{\beta_t, \beta_{t^\star_{\mathsf{unif}}}\} + D\Lambda \rceil$, where D is a sufficiently large constant; decreasing $c_{A,K,L}$ if necessary, $1 \leqslant r \leqslant d - 1$. Write $v = r/d$. Let N_r be the number of elements of $\mathcal{C}$ at Hamming distance exactly r from x. Its expectation is

$$\mu_r = \mathbb{E}_{\mathcal{C}} N_r = M 2^{-d} \binom{d}{r} \,.$$

By Eq. (37) of Lemma F.7 and $\log M = d\,\mathsf{I}_{\mathsf{unif}}(t^\star_{\mathsf{unif}}) = d(\log 2 - h_2(\beta_{t^\star_{\mathsf{unif}}}))$ (Eq. (22)),

$$\log \mu_r \geq d\big(h_2(v) - h_2(\beta_{t^\star_{\mathsf{unif}}})\big) - \log(d+1) \geq cD\Lambda - \log(d+1)\,.$$

The last step uses that $v - \beta_{t^\star_{\mathsf{unif}}} \geq D\Lambda/d$ and that h_2' is lower bounded away from zero in the relevant interval. Since $\Lambda \geq \log d$, choosing D large enough gives $\mu_r \geq 4e^{K\Lambda}$.

Define the Hamming sphere $\mathcal{S}_r(x) = \{z \in \mathcal{X} \mid d_H(x,z) = r\}$. For $z \in \mathcal{S}_r(x)$, denote the indicator that z is included in $\mathcal{C}$ by $\mathbf{I}_z = \mathbb{1}[z \in \mathcal{C}]$. Then $N_r = \sum_{z\in\mathcal{S}_r(x)} \mathbf{I}_z$. For distinct $z, z' \in \mathcal{S}_r(x)$, the events $\{z \in \mathcal{C}\}$ and $\{z' \in \mathcal{C}\}$ are negatively correlated, so $\mathbb{V}_{\mathcal{C}}[N_r] \leqslant \sum_{z\in\mathcal{S}_r(x)} \mathbb{V}_{\mathcal{C}}[\mathbf{I}_z] \leqslant \mu_r$. By Chebyshev's inequality,

$$\mathop{\mathbb{P}}_{\mathcal{C}}[N_r < \mu_r/2] \leqslant 4/\mu_r \leqslant e^{-K\Lambda}\,. \tag{133}$$

Henceforth condition on the event $N_r \geq \mu_r/2$. Each point $z \in \mathcal{S}_r(x) \cap \mathcal{C}$ has the same transition probability $(K_t^{\mathsf{U}})^{\otimes d}(x \mid z)$. Their contribution alone yields

$$\begin{aligned} L_t(x) = \frac{q_t^{\mathcal{C}}(x)}{2^{-d}} &\geq \frac{N_r}{M} 2^d \beta_t^r (1-\beta_t)^{d-r} \\ &\geq \frac{1}{2}\binom{d}{r}\beta_t^r(1-\beta_t)^{d-r} = \frac{1}{2}\,\mathbb{P}[\mathrm{Bin}(d,\beta_t) = r]\,. \end{aligned} \tag{134}$$

The choice of r gives $0 \leqslant v - \beta_t \leqslant |\beta_t - \beta_{t^\star_{\mathsf{unif}}}| + D\Lambda/d + 1/d$. Since $|t - t^\star_{\mathsf{unif}}| \leqslant Lw$ and β_t' is bounded, $|\beta_t - \beta_{t^\star_{\mathsf{unif}}}| \leqslant Cw$. Thus

$$d(v-\beta_t)^2 \leqslant C\left(dw^2 + \frac{D^2\Lambda^2}{d} + \frac{1}{d}\right) \leqslant C'\Lambda\,.$$

Eq. (38) now lower bounds the right-hand side of Eq. (134) by $e^{-C\Lambda}$, the factor $\frac{1}{2(d+1)}$ being absorbed using $\Lambda \geq \log d$. Together with Eq. (133), this proves the lemma. □

J.4.6 Proof of Proposition F.10

Proof. Recall Algorithm 1: the grid $\mathcal{T}$ covers $[\tau_- - 3w, \tau_+ + 3w]$ with mesh $w/20$, each grid point t is tested with $R = 2\lceil C\Lambda\rceil + 1$ independent samples $X^{(j)} \sim \mathrm{Unif}(\mathcal{X})$, and the search stops at the first grid point at which the median of $\widehat{T}_t(X^{(j)})$ over these samples falls below $-(a+b)dw/2$, where $a < b$ are the constants of Proposition F.9.

The search only needs two kinds of decisions to be reliable: it must continue at every grid point above $t^\star_{\mathsf{unif}} + 2w$, and it must stop by the time it has passed through $[t^\star_{\mathsf{unif}} - 3w, t^\star_{\mathsf{unif}} - 2w]$. Proposition F.9 controls a single query in exactly these two regions.

For a relevant grid point t ($t > t^\star_{\mathsf{unif}} + 2w$ or $t^\star_{\mathsf{unif}} - 3w < t < t^\star_{\mathsf{unif}} - 2w$), let $b_{\mathcal{C},t}$ be the probability, conditional on $\mathcal{C}$, that one trial lies on the wrong side of the threshold $-(a+b)dw/2$. Proposition F.9 with exponent 12 gives $\mathbb{E}_{\mathcal{C}} b_{\mathcal{C},t} \leqslant e^{-12\Lambda}$ and hence $\mathbb{P}_{\mathcal{C}}[b_{\mathcal{C},t} > 1/16] \leqslant 16e^{-12\Lambda}$. The grid has $O(1/w) \lesssim e^{\Lambda}$ points for sufficiently large d. A union bound shows that, outside an event over $\mathcal{C}$ of probability at most $16e^{-11\Lambda} \leqslant \delta/4$, every relevant trial has conditional error probability at most $1/16$.

Fix such an $\mathcal{C}$. For each grid point the R trials are independent. The probability that at least half are wrong is at most $2^R(1/16)^{R/2} \leqslant 2^{-R}$. Taking a union bound over the grid and choosing the constant C in the expression of R sufficiently large makes the probability that there is a grid point where at least half of the R trials are wrong at most $\varepsilon/8$. We may draw the randomness for all grid points in advance, so that stopping early does not change this simultaneous guarantee. On the event that more than half of the R trials are correct at all relevant grid points, the search never stops above $t^\star_{\mathsf{unif}} + 2w$, and the grid mesh ensures that it tests a point in $[t^\star_{\mathsf{unif}} - 3w, t^\star_{\mathsf{unif}} - 2w]$, where it must stop, if it has not already done so. This proves Eq. (46). There are $O(1/w)$ grid points and $O(\Lambda)$ queries at each, resulting in the query bound. □

J.4.7 Proof of Lemma F.11

Proof. Write $u = e^{-s}$ and $v = e^{-h}$, so that $e^{-t} = uv$. Substitution of the two clipping endpoints into Eq. (47) gives

$$p_h(\ell_t) = \frac{(1-v)(1-u)}{2(1+uv)}, \qquad p_h(\ell_t^{-1}) = \frac{(1-v)(1+u)}{2(1-uv)}.$$

Because $s, t \in I_0$, the factors other than $1-v$ are bounded above and below by positive constants. Also $1 - v \asymp h$ for $h \leqslant h_0$. Since p_h is increasing, every true or clipped approximate score gives a probability in $[ch, Ch]$. Choose h_0 small enough that the upper bound is at most $1/2$.

Let r and $\widetilde{r}$ be the true and clipped approximate score entries for one coordinate, and write $p = p_h(r)$ and $\widehat{p} = p_h(\widetilde{r})$. The slope of p_h is at most Ch, so $|p - \widehat{p}| \leqslant Ch|r - \widetilde{r}|$. Eq. (39) now gives

$$\mathsf{KL}(\mathrm{Ber}(p) \,\|\, \mathrm{Ber}(\widehat{p})) \leqslant \frac{(p-\widehat{p})^2}{\widehat{p}(1-\widehat{p})} \leqslant Ch(r-\widetilde{r})^2.$$

Summing over the coordinates and applying Eq. (36) proves Eq. (49). □

J.5 Proofs from Section G

Throughout this subsection we use the notation of Section G: Λ is as in Eq. (54) and $w = A\sqrt{\Lambda/d}$ as in Proposition G.4; I_0 is the interval of Eq. (23); ν_t is the reference law of Eq. (56) and $L_t = q_t^{\mathcal{C}}/\nu_t$ the density ratio of Eq. (64); $\lambda_t = e^{-t}/\sigma_t$, p_u, and ϕ_d are as in Eqs. (57) and (58); c_I, C_I are the constants of Eq. (60); A_z, B_z are the log averages of Eq. (67), and $\pi_r(\cdot \mid z)$ and $\mathcal{I}_z$ are the cube posterior and posterior information of Eqs. (69) and (70); m_t, m_t^0, and T_t are as in Eqs. (65) and (66), and $\widetilde{m}_t$, $\widetilde{s}_t$, $\widehat{T}_t$ are their clipped versions from Eqs. (79) and (80); finally Σ_t and V are the posterior covariance and its expected trace from Eq. (87).

J.5.1 Proof of Lemma G.5

Proof. Rather than comparing the mixture laws p_u and p_v directly, we lift them to the joint laws of the latent variable and observation (U, Z). Let P_r denote the joint law

$$U \sim \mathrm{Unif}(\mathcal{X}), \qquad Z = rU + G,$$

for $r > 0$. Conditional on $U = y$, the laws of Z under P_u and P_v are Gaussians with means uy and vy and common covariance I_d. Hence

$$\frac{\mathrm{d}P_u}{\mathrm{d}P_v}(U, Z) = \frac{p_u(Z \mid U)}{p_v(Z \mid U)} = \frac{\phi_d(Z - uU)}{\phi_d(Z - vU)}.$$

Since $\|y\|_2^2 = d$ for every $y \in \mathcal{X}$, the squared distance between these two means is $d(u-v)^2$. Therefore, the second moment of the density ratio under P_v is

$$\mathbb{E}_{P_v}\left[\left(\frac{\mathrm{d}P_u}{\mathrm{d}P_v}\right)^2\right] = \mathbb{E}_{Y \sim \mathrm{Unif}(\mathcal{X})} \int \frac{\phi_d(z - uY)^2}{\phi_d(z - vY)}\,\mathrm{d}z = e^{d(u-v)^2}.$$

Thus, by change of measure and Cauchy–Schwarz,

$$\begin{aligned} p_u(A) &= \mathbb{E}_{P_v}\left[\frac{\mathrm{d}P_u}{\mathrm{d}P_v}(U, Z)\,\mathbb{1}[Z \in A]\right] \\ &\leqslant \left(\mathbb{E}_{P_v}\left[\left(\frac{\mathrm{d}P_u}{\mathrm{d}P_v}\right)^2\right]\right)^{1/2} P_v(Z \in A)^{1/2} = e^{d(u-v)^2/2}\sqrt{p_v(A)}, \end{aligned}$$

proving Eq. (71).

Since $\mathcal{C}$ has the same independent distribution under both laws, adjoining it to the two joint laws does not change the likelihood ratio or its second moment. The same argument therefore proves Eq. (72). □

J.5.2 Proof of Lemma G.6

Proof. The coordinates of Z are independent mixtures of Gaussians with bounded means. Thus Eq. (73) holds except with probability e^{-cd}. Since

$$B_Z''(r) = \sum_{i=1}^{d} Z_i^2 \operatorname{sech}^2(rZ_i)\,,$$

this also gives the upper bound for every r. For the lower bound, count the coordinates for which $1 \leqslant |Z_i| \leqslant 2$. Each coordinate has a probability bounded away from zero of satisfying this condition, uniformly in t. A Chernoff bound shows that at least cd coordinates satisfy it except with probability e^{-cd}. Each such coordinate contributes a fixed positive amount to $B_Z''(r)$ throughout the prescribed interval.

By Eq. (70), $\mathcal{I}_Z(\lambda_t)$ is a sum of d independent one-coordinate $\lambda_t Z_i \tanh(\lambda_t Z_i) - \log\cosh(\lambda_t Z_i)$ terms. Each lies in $[0, \log 2]$, and its mean is $\mathsf{I}_{\mathsf{gauss}}(t)$. Hoeffding's inequality gives Eq. (75). Choose C_K so that its failure probability is at most $e^{-(K+2)\Lambda}$, and then choose c_K small enough to absorb the two probabilities e^{-cd}. □

J.5.3 Proof of Lemma G.8

Proof. Every coordinate of $m_t(x)$ lies in $[-1, 1]$. Tweedie's formula and the fact that clipping cannot increase Euclidean distance to this cube therefore give

$$\|\widetilde{m}_t(x) - m_t(x)\|_2 \leqslant e^t \sigma_t^2 \|\widehat{s}_t(x) - s_t(x)\|_2\,.$$

Multiply this inequality by $e^{-t}\sigma_t^{-2}$ to obtain the corresponding bound for $\widetilde{s}_t - s_t$. Squaring, averaging under $q_t^{\mathcal{C}}$, and using the definition of $\varepsilon_{\mathsf{gauss}}$ proves Eq. (81). Eq. (82) follows by Cauchy–Schwarz. □

J.5.4 Proof of Lemma G.9

Proof. For $u = \lambda_t$ and $z = x/\sigma_t$, Eq. (68) writes the density ratio as

$$L_t(x) = \frac{q_t^{\mathcal{C}}(x)}{\nu_t(x)} = \frac{1}{M}\sum_{y \in \mathcal{C}} e^{u\langle z, y\rangle - B_z(u)}\,. \tag{135}$$

Define the summand

$$L_{u,z}(y) = e^{u\langle z,y\rangle - B_z(u)} = \frac{\phi_d(z - uy)}{p_u(z)} = \frac{\pi_u(y \mid z)}{2^{-d}}\,.$$

Thus $L_{u,z}(y)$ is the posterior-to-prior ratio of y under the whole-cube prior, and

$$\mathbb{E}_{Y \sim \mathrm{Unif}(\mathcal{X})} L_{u,z}(Y) = 1\,.$$

Eq. (135) is therefore the empirical average of these likelihood ratios over the M points of $\mathcal{C}$, sampled without replacement from the cube. We will show that enough moderate likelihood ratios remain in this average to prevent it from being too small.

Fix t in the stated range, and write $u = \lambda_t$ and $Z = X/\sigma_t$. Since $|t - t^\star_{\mathsf{gauss}}| \leqslant Lw$ and $t^\star_{\mathsf{gauss}} \in [\tau_-, \tau_+]$, we have $t \in I_0$ as soon as $Lw \leqslant \tau_-/2$, which holds after decreasing $c_{A,L,K}$; we may therefore apply Lemma G.6 with exponent $K+3$. Since $|t - t^\star_{\mathsf{gauss}}| \leqslant Lw$, Eqs. (60) and (75) imply

$$\mathcal{I}_Z(u) \leqslant d\,\mathsf{I}_{\mathsf{gauss}}(t) + C_{K+3}\sqrt{d\Lambda} \leqslant \log M + (C_I L A + C_{K+3})\sqrt{d\Lambda} = \log M + C_0\sqrt{d\Lambda}\,,$$

where C_0 depends only on A, L, K and $[\kappa_-, \kappa_+]$. Fix an observation z for which this bound and the uniform bounds $cd \leqslant B_z''(r) \leqslant Cd$ hold. Until the last step, all probabilities are conditional on this z.

Now we migrate to a slightly larger noise level. Set

$$v = u - D_0\sqrt{\Lambda/d}\,,$$

where D_0 will be chosen sufficiently large. Since $\mathcal{I}_z'(r) = rB_z''(r) \geq cd$ throughout the relevant parameter interval, decreasing the signal strength from u to v decreases the posterior information $\mathcal{I}_z$ by at least $cD_0\sqrt{d\Lambda}$. Thus, choosing D_0 sufficiently large compared with C_0 gives

$$\mathcal{I}_z(v) \leqslant \log M - 4\sqrt{d\Lambda}\,. \tag{136}$$

Decreasing $c_{A,L,K}$ if necessary ensures that v remains in the same fixed parameter interval.

Draw $Y \sim \pi_v(\cdot \mid z)$. The logarithm of the summand $L_{u,z}(Y)$ has mean

$$\begin{aligned}\mathbb{E}_{Y\sim\pi_v(\cdot|z)} \log L_{u,z}(Y) &= \mathbb{E}_{Y\sim\pi_v(\cdot|z)} \left[u\langle z, Y\rangle - B_z(u)\right] \\ &= \mathcal{I}_z(v) - \mathsf{KL}(\pi_v(\cdot \mid z) \parallel \pi_u(\cdot \mid z)) \leqslant \mathcal{I}_z(v)\,,\end{aligned} \tag{137}$$

and variance

$$\mathop{\mathbb{V}}_{\pi_v}[\log L_{u,z}(Y)] = u^2 B_z''(v) \leqslant Cd\,.$$

Thus, under the slightly noisier posterior, the log likelihood ratio is typically well below $\log M$.

Fix a constant $D > 0$, to be chosen later, and define

$$\mathcal{G} = \left\{y \in \mathcal{X} : L_{u,z}(y) \leqslant Me^{-D\Lambda}\right\}\,.$$

Take $c_{A,L,K}$ sufficiently small that $D\Lambda \leqslant \sqrt{d\Lambda}$. By Eqs. (136) and (137), the cutoff $\log M - D\Lambda$ lies at least $3\sqrt{d\Lambda}$ above the mean of $\log L_{u,z}(Y)$. Chebyshev's inequality therefore gives

$$\pi_v(\mathcal{G} \mid z) \geq 1 - \frac{C}{\Lambda} \geq \frac{1}{2}$$

for sufficiently large d.

We next show that $\mathcal{G}$ still has nonnegligible posterior mass at the original signal strength u. Taylor's theorem and the bound $B_z''(r) \leqslant Cd$ imply

$$\mathsf{KL}(\pi_v(\cdot \mid z) \parallel \pi_u(\cdot \mid z)) = B_z(u) - B_z(v) - (u-v)B_z'(v) \leqslant Cd(u-v)^2 \leqslant CD_0^2\Lambda\,. \tag{138}$$

On the other hand, the data processing inequality for KL divergence gives

$$\begin{aligned}\mathsf{KL}(\pi_v(\cdot \mid z) \parallel \pi_u(\cdot \mid z)) &\geq \mathsf{KL}(\mathrm{Ber}(\pi_v(\mathcal{G} \mid z)) \parallel \mathrm{Ber}(\pi_u(\mathcal{G} \mid z))) \\ &\geq \pi_v(\mathcal{G} \mid z) \log \frac{1}{\pi_u(\mathcal{G} \mid z)} - \log 2 \\ &\geq \frac{1}{2} \log \frac{1}{\pi_u(\mathcal{G} \mid z)} - \log 2\,.\end{aligned}$$

Combining this with Eq. (138) gives

$$\pi_u(\mathcal{G} \mid z) \geq e^{-C_1\Lambda}\,, \tag{139}$$

where C_1 depends on D_0 but not on D.

Now, consider only the points of $\mathcal{C}$ belonging to $\mathcal{G}$, and define

$$\widetilde{L} = \frac{1}{M} \sum_{y \in \mathcal{C}\cap\mathcal{G}} L_{u,z}(y)\,.$$

Its expectation over $\mathcal{C}$ is

$$\mathbb{E}_{\mathcal{C}}\widetilde{L} = \mathbb{E}_{Y\sim\mathrm{Unif}(\mathcal{X})}\left[L_{u,z}(Y)\,\mathbb{1}[Y \in \mathcal{G}]\right] = \pi_u(\mathcal{G} \mid z)\,.$$

Moreover, $L_{u,z}(y) \leqslant Me^{-D\Lambda}$ on $\mathcal{G}$. The variance bound for sampling without replacement therefore gives

$$\begin{aligned}\mathop{\mathbb{V}}_{\mathcal{C}}[\widetilde{L}] &\leqslant \frac{1}{M} \mathop{\mathbb{V}}_{Y\sim\mathrm{Unif}(\mathcal{X})}[L_{u,z}(Y)\,\mathbb{1}[Y \in \mathcal{G}]] \\ &\leqslant \frac{1}{M}\mathbb{E}_{Y\sim\mathrm{Unif}(\mathcal{X})}\left[L_{u,z}(Y)^2\,\mathbb{1}[Y \in \mathcal{G}]\right] \\ &\leqslant e^{-D\Lambda}\mathbb{E}_{Y\sim\mathrm{Unif}(\mathcal{X})}\left[L_{u,z}(Y)\,\mathbb{1}[Y \in \mathcal{G}]\right] \\ &\leqslant e^{-D\Lambda}\pi_u(\mathcal{G} \mid z)\,.\end{aligned}$$

Hence, by Chebyshev's inequality and Eq. (139),

$$\mathop{\mathbb{P}}_{\mathcal{C}}[\widetilde{L} < \pi_u(\mathcal{G} \mid z)/2] \leqslant \frac{4e^{-D\Lambda}}{\pi_u(\mathcal{G} \mid z)} \leqslant 4e^{-(D-C_1)\Lambda}\,.$$

Choose $D > C_1 + K + 4$. Then, except with probability at most $e^{-(K+2)\Lambda}$ over $\mathcal{C}$,

$$L_t(x) \geq \widetilde{L} \geq \pi_u(\mathcal{G} \mid z)/2 \geq e^{-(C_1+1)\Lambda}\,.$$

Finally, add the probability that the conditions imposed on Z fail and enlarge the constant in the statement. This proves Eq. (83). □

J.5.5 Proof of Proposition G.11

Proof. Recall Algorithm 3: the grid $\mathcal{T}$ covers $[\tau_- - 3w, \tau_+ + 3w]$ with mesh $w/20$, each grid point t is tested with $R = 2\lceil C\Lambda \rceil + 1$ independent samples $X^{(j)} \sim \nu_t$, and the search stops at the first grid point at which the median of $\widehat{T}_t(X^{(j)})$ over these samples falls below $-(a+b)dw/2$, where $a < b$ are the constants of Proposition G.10.

We first control the choice of $\mathcal{C}$, and then the randomness of the search conditional on $\mathcal{C}$.

At a fixed grid point above $t^\star_{\mathsf{gauss}} + 2w$, a wrong decision for one sample means $\widehat{T}_t < -(a+b)dw/2$. In $[t^\star_{\mathsf{gauss}} - 3w, t^\star_{\mathsf{gauss}} - 2w]$, it means the reverse inequality. Proposition G.10, with exponent 12, bounds each of these probabilities by $e^{-12\Lambda}$ after averaging over $\mathcal{C}$.

At any one relevant grid point, Markov's inequality shows that the conditional probability of a wrong decision exceeds $1/16$ for at most a $16e^{-12\Lambda}$ fraction of $\mathcal{C}$. There are $O(1/w) \leqslant e^{\Lambda}$ grid points. A union bound therefore removes at most a $16e^{-11\Lambda} \leqslant \delta/4$ fraction of $\mathcal{C}$, and for every remaining $\mathcal{C}$ the conditional error is at most $1/16$ at every relevant point.

Fix $\mathcal{C}$ satisfying the preceding simultaneous guarantee. At one grid point the R trials are independent. The probability that at least half are wrong is at most $2^R(1/16)^{R/2} \leqslant 2^{-R}$. Taking $R = 2\lceil C\Lambda \rceil + 1$ with a sufficiently large constant C makes the union of these events over the grid have probability at most $\varepsilon/8$.

Now condition on all relevant medians being correct. The search cannot stop above $t^\star_{\mathsf{gauss}} + 2w$. The grid contains a point in $[t^\star_{\mathsf{gauss}} - 3w, t^\star_{\mathsf{gauss}} - 2w]$, at which it must stop if it has not already done so. This gives Eq. (85). We may draw all samples in advance, so early stopping does not alter this argument. Finally, there are $O(1/w)$ grid points and $O(\Lambda)$ score oracle queries at each, giving the query complexity bound. □

J.5.6 Proof of Lemma G.12

Proof. Let $Y \sim q^{\mathcal{C}}$ and observe

$$R_r = rY + B_r, \qquad r \geq 0\,,$$

where B is an independent Brownian motion. For a fixed y, the likelihood of the trajectory $(R_v)_{0 \leqslant v \leqslant r}$ relative to standard Brownian motion is $\exp\big(\langle y, R_r\rangle - \frac{r}{2}\|y\|_2^2\big)$. Hence the conditional probability of $Y = y$ given the observations up to time r is proportional to $q^{\mathcal{C}}(y)\exp\big(\langle y, R_r\rangle - \frac{r}{2}\|y\|_2^2\big)$. In particular, this conditional distribution depends on the observations only through R_r. Define

$$m_r := \mathbb{E}[Y \mid R_r], \qquad \Gamma_r := \mathsf{Cov}(Y \mid R_r)\,.$$

Notice that $R_r/\sqrt{r}$ has the same law as $\sqrt{r}Y + G$, where $G \sim N(0, I_d)$ independently. Hence the posterior at time r is the same as the Gaussian-diffusion posterior at time t when $r = \lambda_t^2 = e^{-2t}/\sigma_t^2$.

Subtract the conditional drift from the observation and set

$$W_r = R_r - \int_0^r m_v\,\mathrm{d}v\,.$$

In the observation filtration $\mathcal{F}_r = \sigma(R_v : 0 \leqslant v \leqslant r)$, this is a continuous martingale with quadratic variation rI_d, hence a Brownian motion. Applying Itô's formula to $\mathbb{P}[Y = y \mid R_r]$, we have

$$\mathrm{d}\,\mathbb{P}[Y = y \mid R_r] = \mathbb{P}[Y = y \mid R_r]\langle y - m_r, \mathrm{d}W_r\rangle\,.$$

By definition of m_r and Γ_r,

$$\mathrm{d}m_r = \sum_{y\in\mathcal{C}} y\,\mathrm{d}\,\mathbb{P}[Y = y \mid R_r] = \left(\sum_{y\in\mathcal{C}} \mathbb{P}[Y = y \mid R_r] y(y - m_r)^\intercal\right)\mathrm{d}W_r = \Gamma_r\,\mathrm{d}W_r\,. \tag{140}$$

Since $\|Y\|_2^2 = d$,

$$\mathbb{E}\,\mathrm{Tr}\,\Gamma_r = d - \mathbb{E}\|m_r\|_2^2\,.$$

On the other hand, applying Itô's formula to $\|m_r\|_2^2$ with Eq. (140) gives

$$\mathrm{d}\|m_r\|_2^2 = 2\langle m_r, \Gamma_r\,\mathrm{d}W_r\rangle + \mathrm{Tr}(\Gamma_r^2)\,\mathrm{d}r\,.$$

Taking expectations therefore yields

$$\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\|m_r\|_2^2 = \mathbb{E}\,\mathrm{Tr}(\Gamma_r^2)\,.$$

Combining the preceding two identities,

$$\frac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}\,\mathrm{Tr}\,\Gamma_r = -\mathbb{E}\,\mathrm{Tr}(\Gamma_r^2)\,.$$

Recall that the Gaussian diffusion posterior corresponds to the preceding observation model with $r = \lambda_t^2 = e^{-2t}/\sigma_t^2$. Since

$$\frac{\mathrm{d}}{\mathrm{d}t}\lambda_t^2 = -2e^{-2t}\sigma_t^{-4}\,,$$

by the chain rule we get the derivative in Eq. (88). Finally, $V(t) = d - \mathbb{E}\|m_t(X_t)\|_2^2$ lies in $[0, d]$. □

J.5.7 Proof of Lemma G.13

Proof. Fix $\mathcal{C}$ throughout. There are two quantities to estimate: the norm of the score itself and the norm of its derivative. Both have direct posterior interpretations.

From $X_t = e^{-t}X_0 + \sigma_t G$, we have $\mathsf{Cov}(G \mid X_t) = \frac{e^{-2t}}{\sigma_t^2}\Sigma_t(X_t)$. From Tweedie's formula, we have $s_t(X_t) = -\sigma_t^{-1}\mathbb{E}[G \mid X_t]$. Since $d = \mathbb{E}\|G\|_2^2 = \mathbb{E}\|\mathbb{E}[G \mid X_t]\|_2^2 + \mathbb{E}\,\mathrm{Tr}\,\mathsf{Cov}(G \mid X_t)$, we obtain

$$\mathbb{E}_{q_t^{\mathcal{C}}}\|s_t\|_2^2 = d\sigma_t^{-2} - e^{-2t}\sigma_t^{-4}V(t) \leqslant d\sigma_t^{-2} \leqslant Cd\,. \tag{141}$$

Differentiating the posterior mean gives $\nabla m_t(x) = e^{-t}\sigma_t^{-2}\Sigma_t(x)$, and hence, by Tweedie's formula,

$$\nabla s_t(x) = -\sigma_t^{-2}I_d + e^{-2t}\sigma_t^{-4}\Sigma_t(x)\,.$$

All coefficients are bounded on I_0. Squaring the two terms separately and using Lemma G.12 yields

$$\mathbb{E}_{q_t^{\mathcal{C}}}\|\nabla s_t\|_{\mathrm{F}}^2 \leqslant Cd + C\mathbb{E}_{q_t^{\mathcal{C}}}\,\mathrm{Tr}(\Sigma_t^2) \leqslant C(d + V'(t))\,. \tag{142}$$

This is where the increase of posterior variance enters the discretization error.

The density $q_t^{\mathcal{C}}$ solves $\partial_t q_t^{\mathcal{C}} = \Delta q_t^{\mathcal{C}} + \nabla\cdot(xq_t^{\mathcal{C}})$. Differentiating its logarithm gives

$$\partial_t s_t = \Delta s_t + 2(\nabla s_t)s_t + s_t + (\nabla s_t)x\,.$$

We now apply Itô's formula to $s_{r-u}(Y_u)$, where $\mathrm{d}Y_u = (Y_u + 2s_{r-u}(Y_u))\,\mathrm{d}u + \sqrt{2}\,\mathrm{d}B_u$. The terms containing Δs_t, $(\nabla s_t)s_t$, and $(\nabla s_t)x$ cancel, leaving

$$\mathrm{d}s_{r-u}(Y_u) = -s_{r-u}(Y_u)\,\mathrm{d}u + \sqrt{2}\,\nabla s_{r-u}(Y_u)\,\mathrm{d}B_u\,.$$

Solving this linear equation gives

$$s_{r-u}(Y_u) - s_r(Y_0) = (e^{-u} - 1)s_r(Y_0) + \sqrt{2}\int_0^u e^{-(u-v)}\nabla s_{r-v}(Y_v)\,\mathrm{d}B_v\,.$$

The stochastic integral has conditional expectation zero given Y_0. Eqs. (141)–(142) and Itô's isometry therefore imply

$$\begin{aligned}\mathbb{E}\|s_{r-u}(Y_u) - s_r(Y_0)\|_2^2 &\leqslant Cdu^2 + C\int_0^u (d + V'(r-v))\,\mathrm{d}v \\ &\leqslant Cdu + C(V(r) - V(r-u)), \qquad 0 \leqslant u \leqslant h \leqslant 1\,.\end{aligned}$$

Integrating in u and using the monotonicity of V proves Eq. (89). □

K Details of simulation for Figure 3

Here we describe the numerical simulation that generated Figure 3, which was largely inspired by an experiment carried out for Gaussian diffusion by Xun and Price Xun & Price (2026).

For a fixed codebook $\mathcal{C}$, we consider a planted codeword $y^* \in \mathcal{C}$. Starting from $X_0 = y^*$, we apply the forward process associated to one of the three dLLM paradigms to obtain noisy sample X_t. The posterior distribution given by conditioning on X_t places some mass on y^*, i.e.,

$$\mathbb{P}[X_0 = y^* \mid X_t]\,,$$

and panels (a)-(c) plot empirical estimates for the expectation of this quantity, which we call the *recovery probability*, over the randomness of the forward process and the codebook. In the simulation, we then define the width of the critical window as the distance between the noise levels at which this recovery probability goes from 0.2 to 0.8.

To ensure that the noise levels are normalized the same way, we parametrize by the mutually information between a uniformly random bit and its corrupted version. For uniform and Gaussian diffusion, this information is given by $\mathsf{I}_{\mathsf{unif}}(t)$ and $\mathsf{I}_{\mathsf{gauss}}(t)$, and for masked diffusion with m revealed coordinates, it is $\mathsf{I} = (m/d)\log 2$. The theory then predicts that the critical noise level occurs at $\mathsf{I} = \kappa$, where recall that $\kappa = \frac{1}{d}\log M$.

Because the size of the codebook is exponential in the dimension, it is not possible to directly simulate score access to the random empirical measure. Instead, we use the same Poissonization trick used in Xun & Price (2026): after planting an arbitrary bitstring y^*, we take the remaining "spurious codewords" in the random codebook to be sampled from a Poisson point process on the hypercube with intensity $M2^{-d}$ per point. We elaborate on how to perform an empirical estimate of the recovery probability under this approximation for each of the three frameworks.

Masked diffusion. In this case, for a fixed number of revealed coordinates m, the expected recovery probability is simply the expectation of $\frac{1}{N}$, where N is the number of codewords (including y^*) which agree with y^* on the revealed subset of coordinates. Under Poissonization, $N - 1$ is a draw from $\mathrm{Poi}(M2^{-m})$. Note that the value of m at which the expectation of $\frac{1}{N}$ crosses from 0.2 to 0.8 is clearly independent of d in this setup, so under the normalization by $\mathsf{I} = (m/d)\log 2$, the width of the window in the simulation is $\Theta(1/d)$ as the theory predicts.

Uniform diffusion. In this case, by Lemma B.2,

$$\mathbb{P}[X_0 = y^* \mid X_t] = \frac{1}{1 + \sum_{j=0}^d N_j \tanh(t/2)^{j - d_H(X_t, y^*)}}\,,$$

where N_j is the number of spurious codewords at Hamming distance j from X_t. Note that $d_H(X_t, y^*)$ is distributed as $\mathrm{Bin}(d, \beta_t)$, whereas conditioned on X_t, each quantity N_j is independently distributed as $\mathrm{Poi}(M\binom{d}{j}2^{-d})$. This enables one to empirically estimate the recovery probability, where each sample in the empirical estimate only requires d Poisson samples and one binomial sample.

Gaussian diffusion. In this case,

$$\mathbb{P}[X_0 = y^* \mid X_t] = \frac{1}{1+Z} \qquad \text{for} \qquad Z = \sum_{y \in \mathcal{C} \setminus \{y^*\}} \frac{K_t^{\mathsf{G}}(X_t \mid y)}{K_t^{\mathsf{G}}(X_t \mid y^*)} \,.$$

For spurious codeword y which differs from planted y^* on coordinates $T \subseteq [d]$,

$$\frac{K_t^{\mathsf{G}}(X_t \mid y)}{K_t^{\mathsf{G}}(X_t \mid y^*)} = \exp\Big(-\frac{2e^{-t}}{1-e^{-2t}} \sum_{i \in T} a_i\Big) \qquad a_i = y_i^* X_{t,i} \,.$$

Note that each a_i is independently distributed according to $\mathcal{N}(e^{-t}, 1-e^{-2t})$, and conditional on these draws, T is a uniformly random subset of $[d]$. Under Poissonization, the sum $\sum_{i \in T} a_i$ is thus a Poisson process with intensity $M p_a(s)\,\mathrm{d}s$, where p_a is the density of the random variable $\sum_{i=1}^d b_i a_i$ for $b \sim \mathrm{Unif}(\{0,1\}^d)$. This density is estimated using the saddlepoint method for approximating the pdf of a sum of independent random variables, and the Poisson point process is approximated by discretizing the approximated density into bins.