Proceedings of Machine Learning Research vol 336:1–33, 2026 39th Annual Conference on Learning Theory

Optimal Inference Schedules for Masked Diffusion Models

Sitan Chen SITAN@SEAS.HARVARD.EDU
Harvard University

Kevin Cong KCONG@COLLEGE.HARVARD.EDU
Harvard University

Jerry Li JERRYZLI@CS.WASHINGTON.EDU
UW

Editors: Steve Hanneke and Tor Lattimore

Abstract

A major bottleneck of standard auto-regressive large language models is that their inference process is inherently sequential, resulting in very long and costly inference times. To circumvent this, practitioners proposed a class of language models called *diffusion language models*, of which the *masked diffusion model* (MDM) is the most successful. The MDM is able to sample tokens out-of-order and, ostensibly, many tokens at once and in parallel. However, there is very limited rigorous understanding of how much parallel sampling these models can perform without noticeable degradation in their sampling performance. Prior work in Li and Cai (2025) obtained some preliminary bounds, but these are not tight for many natural classes of distributions. In this work, we give a new, *exact* characterization of the expected divergence between the true distribution and the sampled distribution, for any distribution and any unmasking schedule for the sampler, showing an elegant connection to the theory of *univariate function approximation*.

By leveraging this connection, we then attain a number of novel lower and upper bounds for this problem. While the connection to function approximation in principle gives the optimal unmasking schedule for any distribution, we show that it is in general impossible to compete with it without strong *a priori* knowledge of the distribution, even in seemingly benign settings. However, we also demonstrate new upper bounds and new sampling schedules in terms of well-studied information-theoretic properties of the base distribution, namely, its *total correlation* and *dual total correlation*, which show that in some natural settings, one can sample in $O(\log n)$ steps without any visible loss in performance, where n is the total sequence length.

Keywords: diffusion language models, discrete diffusion, generative modeling, parallel sampling

1. Introduction

Diffusion models are the state-of-the-art approach to generative modeling over domains like video and molecules, and in recent years have also emerged as a powerful alternative (Sahoo et al. (2024); Shi et al. (2024); Nie et al. (2025); Khanna et al. (2025)) to autoregressive large language models (LLMs). Abstractly, these models perform distribution learning by learning to reverse a corruption process transforming data into noise. Starting from fresh noise, they apply the learned reverse process to map it into a fresh sample from the data distribution.

In state-of-the-art diffusion language models, the most common choice of corruption process is the *erasure process*. This is the basis for the popular paradigm of *masked diffusion models (MDMs)*, which now form the backbone of leading approaches to non-autoregressive language modeling. The erasure process proceeds as follows: starting from a sample $X^0 = x^0 \in \Sigma^n$ at time $t = 0$, draw times $T_1, \ldots, T_n$ from some measure on $[0, 1]$, and define $X_i^t = x_i^0$ if $t \leqslant T_i$ and $X_i^t = *$ otherwise, where "$*$" is a special character corresponding to erasure. Given samples from data distribution μ, one then trains a neural network to learn *conditional marginals*: for every time t and conditioning $X^t = x^t$, estimate $\mathrm{law}(X_i^0 \mid X^t = x^t)$ for all $i \in [n]$. It is readily seen that this is equivalent to learning the conditional marginals

$$\mathrm{law}(X_i \mid X_S = z)\,, \qquad S \subseteq [n]\,, i \notin S\,,$$

where $X \sim \mu$, $z \in \Sigma^{|S|}$, and $X_S = z$ denotes the partial assignment to the indices given by S. Given these marginals, it is straightforward to sample from μ by sampling one token at a time. Unlike LLMs which sample one token at a time from left to right, MDMs can sample *out of order*.

Sampling multiple tokens at a time. Crucially in practice, the neural network that is trained to learn these conditional marginals can, given any such partial assignment $X_S = z$, simultaneously compute the conditional marginals for all $i \notin S$ *in one network evaluation*. One of the key selling points of MDMs is thus that these models have the freedom to sample *multiple tokens* at a time in parallel, whereas LLMs are inherently limited to sequential sampling. Empirically, a standard heuristic is the following: fix an *unmasking schedule* given by *step sizes* $s_1, s_2, \ldots, s_k$ summing to n, and iterate the following for $t = 1, \ldots, k$:

- Sample a random subset S_t of size s_t from among the indices $[n] \backslash (S_1 \cup \cdots \cup S_{t-1})$
- For every $i \in S_t$, sample x_i *independently* from $\mathrm{law}(X_i \mid X_{S_1 \cup \cdots \cup S_{t-1}} = x_{S_1 \cup \cdots \cup S_{t-1}})$[1] – note that this ignores correlations across S_t and thus introduces statistical error.

The goal is to make k as small as possible while keeping the statistical error small. If $k = n$ and $s_1 = \cdots = s_n = 1$, then this will perfectly sample from the data distribution μ, but it is no more efficient than sampling with an autoregressive model. On the other hand, if $k = 1$ and $s_1 = n$, this will sample in one step but output the product distribution whose 1-wise marginals agree with μ, which in general will not be a good approximation to μ.

In practice, there is an art to picking the schedule to trade off between these extremes, giving rise to popular heuristics like the *cosine schedule* (Chang et al. (2022); Shi et al. (2024)) and the *log-linear schedule* (Lou et al. (2024); Sahoo et al. (2024)) in which $s_1, s_2, \ldots$ start out small and progressively increase, and more involved adaptive schedules (Kim et al. (2025)). However, our understanding of how to pick these schedules, and how to rigorously quantify the statistical errors that arise from sampling multiple tokens in parallel, remains limited. In this work we therefore ask:

What is the optimal unmasking schedule for a given data distribution μ and target level of error?

This is a challenge not just for theory, but for practice. Indeed, a large-scale ML benchmark Kang et al. (2025) was released just weeks ago in an effort to systematically evaluate unmasking schedules for diffusion language models. But as we will see, this is a question that is particularly amenable to the lens of theory.

1.1. Result 1: Optimal unmasking schedule

Our first result is a tight and surprisingly simple theoretical characterization of the optimal unmasking schedule for any μ. The result exposes an elegant connection to *univariate function approximation*. To state the result, we first require some terminology.

Definition 1 (Expected KL error) *Conditioned on a sequence of subsets $S_1, \ldots, S_k$ of sizes $s_1, \ldots, s_k$, let $\nu^{S_1, \ldots, S_K}$ denote the distribution over outputs x generated by the sampling algorithm above. The notion of error we will consider in this work is the* expected KL error

$$\mathop{\mathbb{E}}_{S_1, \ldots, S_k} [\mathsf{KL}(\mu \,\|\, \nu^{S_1, \ldots, S_k})] ,$$

where the expectation is over subsets S_i of size s_i sampled according to the algorithm above.

1. Of course, in reality we never have exact access to the conditional marginal, but in our theoretical analysis it is straightforward to decouple this error from the overall sampling error; see Appendix G.

Definition 2 (Left Riemann approximation) *Given $\mathbf{Z} = (Z_1, \ldots, Z_n) \in \mathbb{R}^n_{\geq 0}$ and nodes $1 = N_1 < \cdots < N_k < n$, define the* left Riemann approximation *of $\mathbf{Z}$ to be the k-*step sequence $Z_1^{\mathbf{N}}, \ldots, Z_n^{\mathbf{N}}$ *given by:*

$$Z_j^{\mathbf{N}} = \begin{cases} Z_{N_a} & \text{if } N_a \leqslant j < N_{a+1} \\ Z_{N_k} & \text{if } j \geq N_k \end{cases}$$

Given any sequences $\mathbf{Z} = (Z_1, \ldots, Z_n)$ and $\mathbf{Z}' = (Z'_1, \ldots, Z'_n)$, we can define the integration error $\|\mathbf{Z} - \mathbf{Z}'\|_{L^1} \coloneqq \sum_{j=1}^n |Z_j - Z'_j|$. *The k-step left Riemann approximation to $\mathbf{Z}$ minimizing this integration error is:*

$$\mathbf{N}^{*,k} \coloneqq \underset{1=N_1<\cdots<N_k<n}{\operatorname{argmin}} \|\mathbf{Z} - \mathbf{Z}^{\mathbf{N}}\|_{L^1}. \tag{1}$$

Note that given $\mathbf{Z}$, one can efficiently find the minimizing $\mathbf{N}^{*,k}$ in polynomial time via dynamic programming.

The central object in this work is the following sequence quantifying correlations within μ:

Definition 3 (Average mutual information curve) *Given a random variable $X \sim \mu$ over Σ^n, define its* information curve, *denoted $\mathbf{Z} = \mathbf{Z}(\mu)$ by*

$$Z_j = Z_j(\mu) \coloneqq \mathop{\mathbb{E}}_{|S|=j-1, i \notin S}[I(X_i; X_S)], \qquad j \in [n],$$

i.e. the average mutual information between X_i and X_S for random $S \subseteq [n]$ of size $j-1$ and random $i \notin S$.

By Han's inequality (Polyanskiy and Wu, 2025, Theorem 1.7), we have that $0 = Z_1 \leqslant Z_2 \leqslant \cdots Z_n$.

Our first main result is an exact characterization of the optimal expected KL error achievable by any k-step sampler, in terms of the *piecewise approximability of the distribution's information curve*:

Theorem 4 *[Optimal schedule given by best step approximation] Let μ be any distribution over Σ^n, and let $1 \leqslant k \leqslant n$. Let $\mathbf{N}^{*,k}$ be the solution to Eq. (1) for μ's information curve $\mathbf{Z} = \mathbf{Z}(\mu)$.*

Then for any unmasking schedule $s_1, \ldots, s_k$, the expected KL error is given by

$$\mathop{\mathbb{E}}_{S_1,\ldots,S_k}[\mathsf{KL}(\mu \,\|\, \nu^{S_1,\ldots,S_k})] = \|\mathbf{Z} - \mathbf{Z}^{\mathbf{N}}\|_{L^1}, \qquad \text{for} \quad N_a \coloneqq 1 + \sum_{t=1}^{a-1} s_t \quad \forall\, a \in [k].$$

In particular, the schedule that minimizes the expected KL error is

$$s_t = N^{*,k}_{t+1} - N^{*,k}_t, \qquad t \in [k].$$

In Figure 1, we give a pictorial depiction of the expected error in Theorem 4.

The proof of Theorem 4 is remarkably simple once one realizes that the key object driving the statistical error of MDM sampling is the information curve of μ, and we therefore regard the main technical contributions of this result as identifying the correct information-theoretic object to study as well as drawing the surprising connection to univariate function approximation.

1.2. Result 2: Impossibility of competing with the optimal schedule

Although Theorem 4 gives an exact characterization of the optimal schedule, and this schedule can be found in polynomial time given *a priori* knowledge of the information curve, pragmatically it is unclear how to use it as this *a priori* knowledge is not readily available.[2] One might hope that by making use of conditional

2. If one has $\mathrm{poly}(n, \varepsilon)$ many held-out samples from μ, one can always estimate each of the Z_i's to sufficient precision, but in practice it can be prohibitively expensive to generate this many samples using the diffusion model.

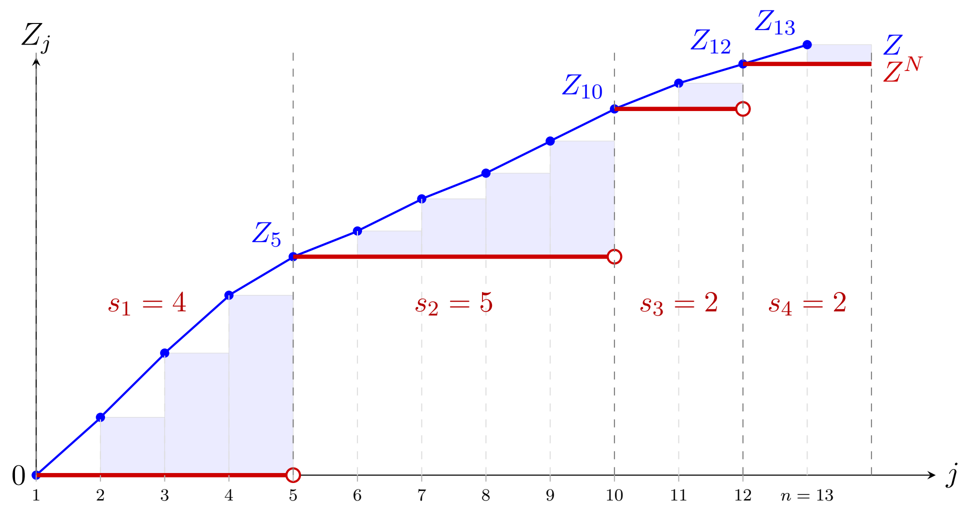

Figure 1: Discrete curve $\mathbf{Z}$ (blue) and left Riemann approximation $\mathbf{Z^N}$ (red) for a sample Z_i curve. The latter extends beyond the Z_j curve to $n+1$ to show the final rectangle $Z_n - Z_{N_{k-1}+1}$; note that this term is not present in a standard left Riemann approximation. Light blue background rectangles represent the Riemann approximation terms. The total area is $\|\mathbf{Z} - \mathbf{Z}^N\|_{L^1}$.

marginal queries to the neural network, one can estimate $\mathbf{Z}(\mu)$ to sufficient accuracy and then deduce the optimal schedule from this.

In the next part of this work, we prove a collection of impossibility results demonstrating that this is not possible in general, even under seemingly benign conditions. Our lower bounds in this part apply to an even more general setting where the sampling algorithm can adaptively make any k conditional marginal queries it chooses (see Definition 11), possibly in a randomized fashion, and then must output a sample such that marginally over its internal randomness, the algorithm's output distribution should be close to μ.

We begin by considering a simple scenario where the curve is promised to either be the constant zero curve ($Z_j = 0$) or a single step function ($Z_j = \mathbb{I}[j > j^*]$ for an unknown j^*), in which case the optimal schedule is simply determined by the location of the step, if it exists. There exist distributions realizing both kinds of information curve, namely the uniform distribution over Σ^n and the uniform distribution over a minimum distance separable (MDS) code (see Definition 17). Our first result shows that even in this situation, finding j^* if it exists requires a prohibitive number of conditional marginal queries.

Theorem 5 (Uniform versus code is hard, see Theorem 24 for formal statement) *There does not exist a single sampling algorithm which simultaneously achieves iteration complexity $o(n)$ for sampling to expected KL error $O(1)$ for all distributions, in fact even for sampling to expected TV error $1/2$. In fact, this holds even if the algorithm knew* a priori *that the distribution were either the uniform distribution over Σ^n or a uniform distribution over an unknown MDS code.*

One might wonder if this worst-case result is too pessimistic, and in practice the relevant data distributions are very far from uniform and have useful correlational structure that one might hope to exploit. Unfortunately, the following strengthening of Theorem 5 shows that this is still not the case:

Theorem 6 (Hardness for general information curves, see Theorem 25 for formal statement) *Let $\mathbf{Z} = \mathbf{Z}(\mu)$ be* any *information curve, where μ is a distribution over $\mathbb{F}_q^n$, such that there exists an unmasking schedule under which one can sample from μ to expected KL error $O(1)$, in fact even just to expected TV error*

$1/2$. *For every* $1 \leqslant k < n$, *let* $\mathbf{Z}^{\uparrow k}$ *denote the information curve given by shifting every* Z_j *for* $j > k$ *up by* $\log_2(q)$.

There does not exist a single sampling algorithm which simultaneously achieves iteration complexity $o(n)$, *even if the algorithm knew* a priori *that the information curve of the distribution was one of* $\mathbf{Z}, \mathbf{Z}^{\uparrow 1}, \ldots, \mathbf{Z}^{\uparrow n-1}$.

Intuitively, this result and the previous one follow from the fact that one can engineer sharp discrete jumps in the information curve which are not detectable unless if one conditions on exactly the right number of indices. We remark that the two results are technically incomparable as the hard distributions for Theorem 6 when $\mathbf{Z}$ is uniformly zero are slightly more intricate than simply uniform vs. MDS.

1.3. Result 3: (Dual) total correlation and reduction to a single hyperparameter sweep

In the final part of this paper, we redeem the situation by showing that for any distribution μ, there exist unmasking schedules depending only on a *single scalar parameter* quantifying correlations in the distribution which achieve small expected KL error.

For this, we need to first define two relevant information-theoretic quantities:

Definition 7 (Total Correlation and Dual Total Correlation) *For any random variable* $X \sim \mu$ *over* Σ^n, *define the* total correlation (TC) *as*

$$\mathrm{TC} = \mathrm{TC}(\mu) := \Big(\sum_{i=1}^{n} H(X_i)\Big) - H(X_1, \ldots, X_n)$$

and the dual total correlation (DTC) *as*

$$\mathrm{DTC} = \mathrm{DTC}(\mu) := H(X_1, \ldots, X_n) - \sum_{i=1}^{n} H(X_i \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)\,.$$

From its definition, we see that TC is equivalently the KL divergence between μ and the product distribution whose marginals agree with those of μ, and thus it characterizes how "product" μ is. On the other hand, DTC has been shown to quantify the extent to which μ can be expressed as a sparse *mixture* of product distributions (Austin (2020)). These quantities admit nice characterizations in terms of the information curve of μ:

Lemma 8 *For any distribution* μ *with information curve* $\mathbf{Z}$,

1. $\mathrm{TC}(\mu) = \sum_{i=1}^{n} Z_i$, *and*
2. $\mathrm{DTC}(\mu) = nZ_n - \sum_{i=1}^{n} Z_i = nZ_n - \mathrm{TC}(\mu)$.

Under the pictorial representation in Figure 1, TC is therefore the area *under* the information curve, and DTC is the area *above* the information curve (capped at Z_n).

We show that while it is not in general possible to compete with a sampler that can choose the unmasking schedule dependent on *a priori* knowledge of the information curve, there are unmasking schedules that only depend on having access to a constant-factor approximations to $\mathrm{TC}(\mu)$ and $\mathrm{DTC}(\mu)$ which only require a number of iterations scaling in $\min(\mathrm{TC}(\mu), \mathrm{DTC}(\mu))$, up to log factors. In situations where these quantities are sublinear in n, this gives us a way to sample asymptotically faster than the naive n-step sampler even without full knowledge of the information curve. As a simple example, if μ is a distribution over a linear subspace of dimension or codimension $O(1)$ (e.g., if μ corresponds to an unknown parity), then this yields an *exponential* speedup over naive schedules. We discuss other such examples in Section 1.4 below.

Theorem 9 (Iteration complexity depending on $\mathrm{TC}, \mathrm{DTC}$) *For any $\varepsilon > 0$, there exists an unmasking schedule which depends only on ε and a parameter $\widehat{\mathrm{TC}}$ (resp. $\widehat{\mathrm{DTC}}$) such that for any distribution μ for which $\mathrm{TC}(\mu) \le \widehat{\mathrm{TC}}$ (resp. $\mathrm{DTC}(\mu) \le \widehat{\mathrm{DTC}}$), the expected KL error satisfies*

$$\mathop{\mathbb{E}}_{S_1,\dots,S_k}[\mathsf{KL}(\mu \,\|\, \nu^{S_1,\dots,S_k})] \leqslant \varepsilon\,,$$

and furthermore the number of steps satisfies

$$k \leqslant 2 + (1+\log n)\cdot(1+\lceil \widehat{\mathrm{TC}}/\varepsilon\rceil) \qquad (\textit{resp. } k \leqslant 2 + (1+\log n)\cdot(1+\lceil \widehat{\mathrm{DTC}}/\varepsilon\rceil)\,.$$

We provide the proof of this in Appendix D; the main idea is to use an exponentially decreasing schedule to attain the TC result and an exponentially increasing schedule to attain the DTC result; therefore, if given both TC and DTC, we can choose the superior schedule. We remark that the latter schedule bears resemblance to what is done in practice, namely unmasking more aggressively as the sampler progresses.

While realizing either (and in particular, the minimum) of these iteration complexities still requires knowing an upper-bound approximation of $\mathrm{TC}(\mu), \mathrm{DTC}(\mu)$, in practice this is not really an issue: one can simply treat these as hyperparameters and either estimate them with held-out data or guess their values via doubling. While the no-go results of Section 1.2 tell us it is impossible in theory to know when to stop doubling, in practice we can simply generate samples according to the different schedules and inspect at what point the output is sufficiently coherent. We emphasize that the reason this is feasible compared to the scheme suggested in Theorem 4 is that we have reduced from designing a schedule that depends on n different hyperparameters $Z_1, \dots, Z_n$ (more than the number of hyperparameters describing the unmasking schedule itself) to designing one that only depends on 2 hyperparameters, namely $\mathrm{TC}(\mu)$ and $\mathrm{DTC}(\mu)$.

Finally, the reader may wonder whether the $\log(n)$ factor in Theorem 9 is a technical artifact or fundamental. In Appendix E we show that it is unavoidable, since there exist information curves which can only be approximated to L^1 error ε using step functions with at least $\Omega(\min(\mathrm{TC}, \mathrm{DTC}) \cdot \log(n)/\varepsilon)$ steps.

1.4. Related work

We contrast our results with some existing bounds from the literature.

The bound of Li and Cai (2025). The most closely related prior work is that of Li and Cai (Li and Cai (2025)). They considered the same setting as the present work and showed that under any unmasking schedule $s_1, \dots, s_k$ with $s_{\max} := \max_i s_i$, the expected KL error can be bounded by

$$\frac{2^{\lceil \log_2 s_{\max}\rceil} - 1}{n} \sum_{i=1}^{n} I(X_i; X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)\,.$$

This was proven using a delicate inductive argument based on recursively relating the expected KL error for a given unmasking schedule to the expected KL error with a schedule whose steps are twice as fine.

We make two observations about this bound. First, armed with Theorem 4, which gives an *exact* characterization of the expected KL error for *any* unmasking schedule, we can give a proof of Li and Cai's bound in just four lines — see Appendix F.1. Second, we note that up to $\log n$ factors, the bound in Theorem 9 is strictly better. The reason is that

$$\sum_{i=1}^{n} I(X_i; X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) = nZ_n = \mathrm{TC}(\mu) + \mathrm{DTC}(\mu) \asymp \max(\mathrm{TC}(\mu), \mathrm{DTC}(\mu))\,.$$

For instance, in the aforementioned simple example where μ is distributed over a generic linear subspace, $\mathrm{TC}(\mu) + \mathrm{DTC}(\mu) = \Theta(n)$, whereas $\min(\mathrm{TC}(\mu), \mathrm{DTC}(\mu))$ scales with the minimum of the dimension and codimension, which can be much smaller (see Example 1).

DTC and the work of Austin (2020, 2019). The elegant work of Austin (2020) gave a powerful operational characterization of DTC. First, it is easily seen that any distribution μ which is a mixture of $2^{o(n)}$ product distributions has $\mathrm{DTC}(\mu) = o(n)$. Austin showed an approximate converse: if $\mathrm{DTC}(\mu) = o(n)$, then μ is well-approximated by a mixture of $2^{o(n)}$ product distributions. In fact, his proof is algorithmic and has an interesting interpretation under the perspective of the present work: it shows that if one first samples $O(\sqrt{\mathrm{DTC}(\mu) \cdot n/\varepsilon})$ indices in sequence and then samples the remaining indices in $\mathcal{O}\left(\sqrt{\mathrm{DTC}(\mu) \cdot n/\varepsilon}\right)$ iterations, one can achieve expected KL error ε:

Theorem 10 (Austin's iteration complexity bound (Austin (2020))) *For any $\varepsilon > 0$, there exists an unmasking schedule which depends only on ε and a parameter $\widehat{\mathrm{DTC}}$ such that for any distribution μ for which $\mathrm{DTC}(\mu) \lesssim \widehat{\mathrm{DTC}}$, the expected KL error satisfies*

$$\mathop{\mathbb{E}}_{S_1,\dots,S_k}[\mathsf{KL}(\mu \,\|\, \nu^{S_1,\dots,S_k})] \leqslant \varepsilon\,,$$

and furthermore the number of steps satisfies $k \lesssim \left\lceil \sqrt{\mathrm{DTC} \cdot n/\varepsilon} \right\rceil$.

We provide a proof for completeness in Appendix F.2. Although this bound is already sublinear in n when $\mathrm{DTC}(\mu) = o(n)$, it is bottlenecked at $\sqrt{n}$. Indeed, note that Austin's bound is the geometric mean of our stronger bound in Theorem 9 and the trivial iteration complexity bound of n, up to a logarithmic factor.

The connection between parallel sampling and decomposition of measure has been quite fruitful within probability theory. For instance, Austin (Austin (2019)) showed the remarkably general result that for any Gibbs measure $\propto e^{-\beta H}$ of "low-complexity" in the sense that the discrete derivatives of the Hamiltonian lie within a set of bounded metric entropy, the DTC is $o(n)$. The decomposition of such distributions into mixtures of product measures arises naturally in the theory of nonlinear large deviations, see e.g., Chatterjee and Dembo (2016); Eldan and Gross (2018a); Eldan (2018); Eldan and Gross (2018b). The theory of parallel sampling can thus also be fruitfully interpreted as providing richer and more accurate hierarchical measure decompositions, where the levels of the hierarchy correspond to iterations of the sampler.

Pinning lemma and stochastic localization. Finally, we note a closely related notion from statistical physics and theoretical computer science, namely the *pinning lemma* (Raghavendra and Tan (2012); Andrea (2008); El Alaoui and Montanari (2022)). The premise behind this result is that if one conditions on a random subset of coordinates of size s according to their true s-wise marginal in μ, then the remaining coordinates have pairwise correlation which is bounded by $O(1/s)$. In some sense this is the fundamental premise behind MDMs: pinning random tokens reduces the correlation among the remaining tokens, which intuitively enables more aggressive parallel sampling of later tokens. This has been used to great effect in the context of SDP rounding algorithms for solving dense CSPs (Raghavendra and Tan (2012); Manurangsi and Raghavendra (2017); Yoshida and Zhou (2014); Jain et al. (2019))

That being said, the pinning lemma holds for *all* distributions, whereas our impossibility results show that without additional prior information about the distribution, one cannot simultaneously achieve $o(n)$ complexity for all μ. It is worth contrasting this state of affairs with the work of Anari et al. (2024). By leveraging the pinning lemma, they showed that in a much *stronger* parallel model where in each round one can simultaneously make multiple conditional marginal queries, each corresponding to a possibly *different* partial assignment, it is possible to sample in $\tilde{O}(n^{2/3})$ parallel rounds, for general distributions.

Other theoretical works on discrete diffusion. We briefly mention some other works in the discrete diffusion literature that derive theoretical bounds. In Chen and Ying (2024), the authors study discretization bounds for a different paradigm of discrete diffusions, where the corruption process being reversed is a bit flip channel rather than an erasure channel. Here, it is nontrivial even to derive bounds which scale linearly in n. In Ren et al. (2025), the authors consider finding better discretizations of the continuous-time Markov chain associated to the discrete diffusion model; under some smoothness assumptions on the underlying distribution, which are primarily relevant to the bit flip setting, their higher-order solvers achieve nontrivial sampling guarantees relative to naive (Euler) discretization.

We also remark that the conditional marginal oracle we consider is very similar in spirit to the *conditional query* model in distribution testing, pioneered by Chakraborty et al. (2013); Canonne et al. (2015). Our model is different in two ways: (1) we restrict to *subcube conditionings* in the sense of Canonne et al. (2021), and (2) a single oracle query gives an entire vector of 1-wise conditional marginals, rather than just a single sample from the posterior. The literature here is extensive and orthogonal to our work; we refer the interested reader to (Canonne, 2020, Chapter 11). Lastly, we note that masked diffusion models – and autoregressive models – can be thought of as modern instantiations of the classical Jerrum-Valiant-Vazirani counting-to-sampling reduction Jerrum et al. (1986).

Continuous diffusion. In recent years there has been significant progress on understanding discretization bounds for diffusion models over continuous spaces, e.g., the works of Chen et al. (2023b); Lee et al. (2023); Chen et al. (2023a); Benton et al. (2024); Conforti et al. (2025); Li et al. (2024). The techniques in this area are largely distinct from the ones in this work, with the exception of the recent work of Reeves and Pfister (2025) which derived an analogous expression to our Theorem 4 for the discretization error incurred by *continuous* diffusions. In that context, the analogue of our information curve is the *MMSE curve* $\mathbb{E}[X - \mathbb{E}[X \mid \alpha_t X + \beta_t \gamma]]$ for Gaussian γ, and they show (see Lemma 2 therein) that the discretization error is exactly given by the left Riemann integration error to this curve. Interestingly, whereas in the continuous diffusion context this integration error exactly characterizes the KL error in *path space*, which is only an upper bound to the KL error at the endpoint of the sampler, in the masked diffusion setting the integration error exactly characterizes the sampler's KL error. In light of the connection to Reeves and Pfister (2025), it would be interesting to extend our impossibility results and TC/DTC-based bounds to the continuous setting.

Concurrent work. Independent concurrent work of Lavenant and Zanella (2025) also identified the connection to Riemann approximation of the information curve (our "Result 1"). Unlike our work, they did not explore the *query complexity* of learning an optimal schedule (our "Result 2") and did not devise explicit schedules that scale better than the bound in Li and Cai (2025) (our "Result 3"). Instead, they additionally provided worst-case bounds for sampling error under arbitrary, non-random orderings, and studied a natural $n \to \infty$ scaling limit of the step function approximation problem.

2. Technical preliminaries

In this section, we will first provide a brief overview of our notation and oracle model. We will then discuss some important information theoretic quantities and results.

2.1. Notation

Throughout the remainder of this paper, we will use the following notation.

Vocabulary, data distribution, and product distributions. We will let Σ be a vocabulary and μ be the data distribution over Σ^n. We use $\Delta(\Sigma)$ to denote the probability simplex over Σ. Let $\mathbf{X} = (X_1, \ldots, X_{|S|}) \sim \mu$. For any set $S \subseteq [n]$, let $X_S = \{X_i\}_{i \in S}$. Define $\mu(\cdot \mid S)$ to be the conditional distribution of μ given S, and $\mu^{\otimes}(\cdot \mid S)$ to be the product distribution which has the same marginals as $\mu(\cdot \mid S)$. Lastly, define $f_\mu(\cdot \mid S)$ and $f_\mu^{\otimes}(\cdot \mid S)$ be the corresponding probability mass functions. Lastly, we set X_S to denote the set $\{X_i\}_{i \in S}$.

2.2. Oracle model

Throughout this work, our main oracle object will be the *conditional marginal oracle*, which outputs the marginals of μ conditioned on any subset. To define it, first recall that $(X_1, \ldots, X_n) \sim \mu$ is the data distribution over a vocabulary Σ. Let $\mathcal{D}$ be the collection of all multivariate distributions Σ. Moreover, let $\mathbf{p}_{i|S}(x_S) = \{p(X_i = j \mid X_S = x_S), j \in \Sigma\}$ be the marginal probability vector on coordinate i. We then have the following.

Definition 11 (Conditional marginal oracle) *The* conditional marginal oracle CO *takes as input a partial assignment $X_S = s$ and outputs the conditional marginal distributions of μ given $X_S = x_S$. Formally,*

$$\mathsf{CO}(X_i \mid X_S = x_S) = \{\mu(X_i \mid X_S = x_S)\}_{i \notin S}.$$

If the pinning $X_S = x_S$ is impossible in $\operatorname{supp}(\mu)$*, output an arbitrary element of* $\Delta(\Sigma)^{n-|S|}$.

In our upper bounds, we will only ever use the oracle to obtain conditional marginals to sample from in parallel, as this is the standard way in practice to use this oracle. Our lower bounds however apply to the most general setting of arbitrary randomized algorithms with adaptive query access to CO (see Definition 22).

Note that CO is an exact oracle, whereas in practice, an approximate oracle $\widehat{\mathsf{CO}}$ is learned from the training data. However, in Appendix G, we show following Li and Cai (2025) that the error of our sampling algorithms can be decoupled into learning and sampling error. Since this work focuses on the sampling procedure and error, we will assume that the learned oracle is perfect.

2.3. Information-theoretic quantities

In this section, we will recall some information theoretic quantities and prove a few preliminary lemmas which will be useful in the subsequent proofs of our main results. First, recall that $H(X)$ refers to the entropy of a random variable, and $H(Y \mid X)$ refers to the conditional entropy of a pair of random variables.

Moreover, recall the definitions of TC and DTC from Definition 7. To provide some intuition for these quantities, we first provide some examples of the TC and DTC of linear subspace and product mixture distributions.

Example 1 (Linear Subspaces) *Suppose $\Sigma = \mathbb{F}_q$ and $\mathcal{V} \subseteq \Sigma^n$ denote a linear subspace of dimension d. Let $\mu_{\mathcal{V}}$ denote the uniform distribution over $\mathcal{V}$. Then there is a matrix M such that $MU \sim \mu_{\mathcal{V}}$ for a uniform $U \in \mathbb{F}_q^d$. Then*

$$\mathrm{TC}(\mu) = (n-k)\log q - d\log q = (n-d-k)\log q$$

where $k = \#\{i : M_i = 0\}$ is the number of rows of M which are identically 0. Moreover,

$$\mathrm{DTC}(\mu) = d\log q - \ell\log q = (d-\ell)\log q,$$

where $\ell = \#\{i : M_i \notin \operatorname{span}(\{M_j\}_{j\neq i})\}$ is the number of rows of M which are not in the span of the remaining rows. In general, we will often have $k = \ell = 0$, in which case $\mathrm{TC}(\mu)$ and $\mathrm{DTC}(\mu)$ are the codimension and dimension of $\mathcal{V}$, respectively.

Example 2 (Mixtures of Products) *The DTC of product mixtures has been studied in-depth before in Austin (2020). In particular, by Proposition 8.1 of Austin (2020), the DTC of a mixture of m product distributions is at most $\log m$. Thus, any mixture μ of $2^{o(n)}$ products satisfies $\mathrm{DTC}(\mu) = o(n)$.*

In the converse direction, by Theorem A of Austin (2020), any μ with $\mathrm{DTC}(\mu) = o(n)$ is close in transport distance to a relatively "simple" mixture of products.

Next, we define a sequence of values based on the average entropy of fixed-cardinality subsets. They will be useful to help analyze the information curve.

Definition 12 (Average entropy curve) *The* average entropy curve *of the distribution* $(X_1, \ldots, X_n) \sim \mu$ *is given by*[3]

$$H_i(\mu) = \frac{1}{\binom{n}{i}} \sum_{S \subseteq [n], |S|=i} H\left(\{X_i\}_{i \in S}\right) = \mathbb{E}_{|S|=i} H\left(\{X_j\}_{j \in S}\right) .$$

When μ *is clear from context, we denote this by* H_i.

We can then express the information curve in terms of the average entropy curve as follows.

Lemma 13 *We have* $Z_i = H_1 + H_{i-1} - H_i$.

The proof of this is presented in Appendix A.

3. Sampling error in terms of unmasking schedule: proof of Theorem 4

In this section we establish an upper bound for the expected KL error of a fixed and then random unmasking algorithm, effectively proving Theorem 4. We first formalize the definition of fixed unmasking algorithms as follows.

Definition 14 (Fixed Unmasking Algorithm) *The fixed unmasking algorithm with subset schedule* $(S_1, \ldots, S_k)$, *given by* $\mathcal{A}_{\text{fixed}}(k, \{S_i\}_{i=1}^k)$, *proceeds as follows. First define* $N_i = \sum_{j=1}^i s_j$, *where* $N_0 = 0$ *and* $s_j = |S_j|$. *Then at each stage* $i \in [k]$, *beginning at* $i = 1$, *independently and in parallel sample*

$$x_j \sim \mu\left(X_j \mid X_{\bigsqcup_{t=1}^{i-1} S_t}\right)$$

for all $j \in S_i$. *The algorithm then outputs the sample* $(x_1, \ldots, x_n) \sim \nu^{S_1, \ldots, S_k}$.

We next define the random unmasking algorithm as follows, based on the fixed unmasking algorithm. This is the formal definition for the algorithm that was outlined in Section 1.

Definition 15 (Random Unmasking Algorithm) *The* random unmasking algorithm with unmasking schedule $(k, \{s_i\}_1^k)$, *given by* $\mathcal{A}(k, \{s_i\}_1^k)$, *proceeds as follows. First, sample a uniformly random partition of coordinates* $S = \sqcup_{i=1}^k S_i, |S_i| = s_i$. *Output a sample* $(x_1, \ldots, x_n) \sim v^{S_1, \ldots, S_k}$ *given by* $\mathcal{A}_{\text{fixed}}(k, \{S_i\}_{i=1}^k)$. *The algorithm then outputs the sample* $(x_1, \ldots, x_n) \sim \nu$.

Note that the fixed unmasking algorithm $\mathcal{A}_{\text{fixed}}$ essentially chooses s_i tokens at fixed positions at each stage i and samples all tokens independently and in parallel, and the random unmasking algorithm selects the token positions at each stage uniformly at random amongst all masked tokens. We can now formally state and prove (a slightly stronger form of) Theorem 4.

Theorem 16 *Let* μ *denote the underlying distribution of data. Let* $(k, \{s_i\}_1^k)$ *be an unmasking schedule and* $\{S_i\}_1^k, |S_i| = s_i$ *be a fixed subset schedule. Let* $N_i = \sum_{j=1}^i s_j$ *denote the partial sums of the* s_i *sequence, where* $N_0 = 0$. *Suppose the fixed unmasking algorithm* $\mathcal{A}_{\text{fixed}}(\{S_i\}_1^k)$ *samples a distribution* $\nu^{S_1, \ldots, S_k}$ *and the random unmasking algorithm* $\mathcal{A}(k, \{s_i\}_1^k)$ *samples a distribution* ν. *Then both algorithms have query complexity* k, *and* ν *achieves KL error relative to* μ *of*

$$\mathsf{KL}(\mu \,\|\, \nu) \le \mathbb{E}_{S_1, \ldots, S_k}\left[\mathsf{KL}\left(\mu \,\middle\|\, \nu^{S_1, \ldots, S_k}\right)\right] = \sum_{i=1}^k \left(\sum_{j=1}^{s_i} \left(Z_{N_{i-1}+j} - Z_{N_{i-1}+1}\right)\right) ,$$

where the expectation is taken over all partitions $S = \bigsqcup_{i=1}^k S_i$ *for which* $|S_i| = s_i$.

3. For $i = 0$, the entropy of the empty set is defined to be 0.

The proof of this result is in Appendix B. We now make two brief comments about Theorem 16.

Theorem 16 and Theorem 4. First, we note that Theorem 4 is an immediate corollary; see Appendix B.

Comparison between fixed and random unmasking algorithm. There are two methods of approaching sampling: first, by fixing the schedule S_i ahead of time, and second, by resampling S_i from $|S_i| = s_i$ for each sample. These correspond to the fixed and random unmasking algorithms, respectively. We observe that the inequality in Theorem 16 shows that the distribution outputted by the *random* unmasking algorithm is, on average, at least as good in terms of KL-error from μ to the distribution outputted by the *fixed* unmasking algorithm. This is an additional guarantee not given in Theorem 4, and suggests that the random unmasking algorithm is superior, albeit requiring an additional step in the sampling process.

4. Lower bounds on competing with the oracle rate

Definition 17 (MDS codes) *A k-dimensional linear subspace $\mathcal{V}$ of $\mathbb{F}_q^n$ is an* maximum distance separable (MDS) code *if for any $k \times n$ matrix M whose rows constitute a basis for $\mathcal{V}$, every k columns of M are linearly independent. We denote by $\mathrm{Unif}(\mathcal{V})$ the uniform distribution over points in $\mathcal{V}$.*

In this work, we will consider affine shifts *of MDS codes. That is, we will consider distributions over* affine *subspaces which are given by taking some MDS code and translating it by a fixed vector in $\mathbb{F}_q^n$. We will abuse terminology and refer to such affine subspaces as MDS codes.*

We will consider "random" MDS codes:

Definition 18 (Balanced random MDS codes) *A distribution $\mathcal{D}$ over k-dimensional MDS codes is* balanced *if for every subset $S \subseteq [n]$ for which $|S| \geq k$ and every partial assignment $x \in \mathbb{F}_q^{|S|}$,*

$$\mathop{\mathbb{P}}_{\mathcal{V}\sim\mathcal{D}}[\exists\, x^* \in \mathcal{V} : x^*_S = x] = (1/q)^{|S|-k}$$

Reed-Solomon codes provide an example of MDS codes. Below, we recall their definition.

Definition 19 (Reed-Solomon codes) *Let q be any prime power exceeding n, and let k be any value between 1 and $n-1$. A k-*dimensional Reed-Solomon (RS) code *in $\mathbb{F}_q^n$ is a linear subspace specified as follows. It is specified by a collection of distinct* evaluation points *$a_1, \ldots, a_n \in \mathbb{F}_q$, and is given by the set of all evaluations $(p(a_1), \ldots, p(a_n))$ where p is a polynomial over $\mathbb{F}_q$ of degree less than k.*

As in Definition 17, we will abuse terminology and also refer to affine shifts of RS codes as RS codes.

We will leverage the following basic property of MDS codes:

Proposition 20 *Let $\mu = \mathrm{Unif}(\mathcal{V})$ for any k-dimensional MDS code $\mathcal{V} \subseteq \mathbb{F}_q^n$. Then for any $S \subseteq [n]$ satisfying $|S| < k$ and any partial assignment $x \in \mathbb{F}_q^{|S|}$, $\mu(X_i \mid X_S = x) = \mathrm{Unif}(\mathbb{F}_q)$ for all $i \notin S$.*

In particular, this implies that $Z_j(\mu) = \log_2(q) \cdot \mathbb{I}[j > k]$.

In addition, recall from the definition of the oracle that if $|S| > k$ and the partial assignment $X_S = x$ is incompatible with any element of $\mathcal{V}$, then the output of the conditional marginal oracle can be arbitrary. Throughout this section, we will take the oracle's output in this case to be $\mathrm{Unif}(\mathbb{F}_q)^{\otimes(n-|S|)}$.

We will also use the following property of *random* Reed-Solomon codes. Given prime power $q \geq n$ and dimension $0 < k < n$, let $\mathcal{D}_{n,k,q}$ denote the following distribution over k-dimensional RS codes over alphabet q. When n, q are clear from context, we denote this by $\mathcal{D}_k$.

Lemma 21 *$\mathcal{D}_k$ is balanced in the sense of Definition 18.*

Next, we formalize the model of computation under which we prove a lower bound.

Definition 22 (Sampling algorithm) *Let $\mathcal{F} \subseteq \Delta(\Sigma^n)$ denote some known family of distributions μ. Given access to the conditional marginal oracle for some $\mu \in \mathcal{F}$, an $\mathcal{F}$-*aware sampling algorithm *$\mathcal{A}$ is a procedure of the following form:*

1. *Repeat the following:*
   - *Based only on the query outcomes from previous rounds and $\mathcal{F}$ (and not on knowledge of μ), and possibly using additional randomness, either query the oracle on a partial assignment $X_S = x$, or exit the loop.*
   - *If the former, observe conditional marginals $\{\mu(X_i \mid X_S = x)\}_{i \notin S}$*
2. *Output a string in Σ^n.*

*Importantly, the decision to exit out of the loop can be made adaptively. We say that $\mathcal{A}$ is T-*query *if with probability 1 it performs at most T queries to the oracle before terminating. We denote by $\mathcal{A}[\mu]$ the distribution over outputs of $\mathcal{A}$.*

Definition 23 (Query budget and cost function) *Let $\mathcal{T} : \Delta(\Sigma^n) \to \mathbb{N}$ denote a* query budget *for the query complexity of such a sampler, and define*

$$\mathsf{cost}^{\mathsf{KL}}_{\mathcal{T}}(\mathcal{A};\mu) := \begin{cases} \mathsf{KL}(\mu \,\|\, \mathcal{A}[\mu]) & \mathcal{A} \text{ is at most } \mathcal{T}(\mu)\text{-query} \\ \infty & \text{otherwise} \end{cases}$$

Define $\mathsf{cost}^{\mathsf{TV}}_{\mathcal{T}}(\mathcal{A};\mu)$ in the same way, with KL replaced by TV.

Warmup example. We begin by exhibiting a simple ensemble of distributions for which no single algorithm can successfully sample from μ to error ε using $O(\min(\mathrm{TC}(\mu), \mathrm{DTC}(\mu)) \log(n)/\varepsilon)$ for every μ in the ensemble.

Let $\mathcal{U}$ denote the uniform distribution over $\mathbb{F}_q^n$, and let $\mathcal{F}$ consist of $\mathcal{U}$ as well as $\mathcal{U}_{\mathcal{V}}$ for all Reed-Solomon codes $\mathcal{V} \subseteq \mathbb{F}_q^n$ of dimension $0 < k < n$. Formally, we show:

Theorem 24 *No $\mathcal{F}$-aware sampling algorithm $\mathcal{A}$ can achieve $\sup_{\mu \in \mathcal{F}} \mathsf{cost}^{\mathsf{TV}}_{\mathcal{T}}(\mathcal{A};\mu) \leqslant 1/16$ for any budget $\mathcal{T}$ satisfying $\mathcal{T}(\mu) \lesssim \max(1, \min(\mathrm{TC}(\mu), \mathrm{DTC}(\mu))) \log(n)$ for all $\mu \in \mathcal{F}$.*

We prove this in Appendix C.1.

Lower bounds for arbitrary information curves. Although the lower bound of Theorem 24 is quite tailored to distributions over MDS codes, it turns out that the same idea can be extended to show that *any* information curve admits a realization by some distribution which cannot be distinguished from a distribution with the same information curve except shifted upwards by an additive constant for all indices past a certain point.

Theorem 25 *Let $\mathbf{Z}$ be the information curve associated to some distribution with total correlation TC and dual total correlation DTC. Suppose that $\min(\mathrm{TC}, \mathrm{DTC}) \log n \ll n$. Given $1 \leqslant k < n$, let $\mathbf{Z}^{\uparrow k}$ denote the information curve given by*

$$Z_j^{\uparrow k} := \begin{cases} Z_j & \text{if } j \leqslant k \\ Z_j + \log_2(q) & \text{if } j > k \end{cases}$$

There exist a family $\mathcal{F}$ of distributions whose information curves are from among $\{\mathbf{Z}, \mathbf{Z}^{\uparrow 1}, \ldots, \mathbf{Z}^{\uparrow n-1}\}$ such that for every k there is at least one such distribution in $\mathcal{F}$, and furthermore for any budget $\mathcal{T}$ satisfying $\mathcal{T}(\mu) \lesssim \max(1, \min(\mathrm{TC}(\mu), \mathrm{DTC}(\mu))) \log(n)$ for all $\mu \in \mathcal{F}$, no $\mathcal{F}$-aware sampling algorithm $\mathcal{A}$ can achieve $\sup_{\mu \in \mathcal{F}} \mathsf{cost}^{\mathsf{TV}}_{\mathcal{T}}(\mathcal{A};\mu) \leqslant 1/16$.

We prove this in Appendix C.2.

References

Nima Anari, Ruiquan Gao, and Aviad Rubinstein. Parallel sampling via counting. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 537–548, 2024.

Montanari Andrea. Estimating random variables from random sparse observations. *European Transactions on Telecommunications*, 19(4):385–403, 2008.

Tim Austin. The structure of low-complexity gibbs measures on product spaces. *The Annals of Probability*, 47(6):4002–4023, 2019.

Tim Austin. Multi-variate correlation and mixtures of product measures. *Kybernetika*, 56(3):459–499, 2020.

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

Lucien Birgé. Estimating a density under order restrictions: Nonasymptotic minimax risk. *The Annals of Statistics*, pages 995–1012, 1987.

Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020.

Clément L Canonne, Dana Ron, and Rocco A Servedio. Testing probability distributions using conditional samples. *SIAM Journal on Computing*, 44(3):540–616, 2015.

Clément L Canonne, Xi Chen, Gautam Kamath, Amit Levi, and Erik Waingarten. Random restrictions of high dimensional distributions and uniformity testing with subcube conditioning. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 321–336. SIAM, 2021.

Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 561–580, 2013.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022.

Sourav Chatterjee and Amir Dembo. Nonlinear large deviations. *Advances in Mathematics*, 299:396–450, 2016.

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

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

Qi Chen, Chen He, Lingge Jiang, and Qingchuan Wang. Average entropy functions. In *2009 IEEE International Symposium on Information Theory*, pages 2632–2633. IEEE, 2009.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023b.

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

Ahmed El Alaoui and Andrea Montanari. An information-theoretic view of stochastic localization. *IEEE Transactions on Information Theory*, 68(11):7423–7426, 2022.

Ronen Eldan. Gaussian-width gradient complexity, reverse log-sobolev inequalities and nonlinear large deviations. *Geometric and Functional Analysis*, 28(6):1548–1596, 2018.

Ronen Eldan and Renan Gross. Decomposition of mean-field gibbs distributions into product measures. *Electronic Journal of Probability*, 23, 2018a.

Ronen Eldan and Renan Gross. Exponential random graphs behave like mixtures of stochastic block models. *The Annals of Applied Probability*, 28(6):3698–3735, 2018b.

Vishesh Jain, Frederic Koehler, and Andrej Risteski. Mean-field approximation, convex hierarchies, and the optimality of correlation rounding: a unified perspective. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1226–1236, 2019.

Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical computer science*, 43:169–188, 1986.

Wonjun Kang, Kevin Galim, Seunghyuk Oh, Minjae Lee, Yuchen Zeng, Shuibai Zhang, Coleman Hooper, Yuezhou Hu, Hyung Il Koo, Nam Ik Cho, et al. Parallelbench: Understanding the trade-offs of parallel decoding in diffusion llms. *arXiv preprint arXiv:2510.04767*, 2025.

Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, et al. Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298*, 2025.

Seo Hyun Kim, Sunwoo Hong, Hojung Jung, Youngrok Park, and Se-Young Yun. Klass: Kl-guided fast inference in masked diffusion models. *arXiv preprint arXiv:2511.05664*, 2025.

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

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

Gen Li and Changxiao Cai. A convergence theory for diffusion language models: An information-theoretic perspective. *arXiv preprint arXiv:2505.21400*, 2025.

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

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, pages 32819–32848, 2024.

Pasin Manurangsi and Prasad Raghavendra. A birthday repetition theorem and complexity of approximating dense csps. In *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*, pages 78–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017.

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.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

Prasad Raghavendra and Ning Tan. Approximating csps with global cardinality constraints using sdp hierarchies. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 373–387. SIAM, 2012.

Galen Reeves and Henry D Pfister. Information-theoretic proofs for diffusion sampling. *arXiv preprint arXiv:2502.02305*, 2025.

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. *arXiv preprint arXiv:2502.00234*, 2025.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37:103131–103167, 2024.

Yuichi Yoshida and Yuan Zhou. Approximation schemes via sherali-adams hierarchy for dense constraint satisfaction problems and assignment problems. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 423–438, 2014.

Contents

1 Introduction **1**
- 1.1 Result 1: Optimal unmasking schedule . . . 2
- 1.2 Result 2: Impossibility of competing with the optimal schedule . . . 3
- 1.3 Result 3: (Dual) total correlation and reduction to a single hyperparameter sweep . . . 5
- 1.4 Related work . . . 6

2 Technical preliminaries **8**
- 2.1 Notation . . . 8
- 2.2 Oracle model . . . 8
- 2.3 Information-theoretic quantities . . . 9

3 Sampling error in terms of unmasking schedule: proof of Theorem 4 **10**

4 Lower bounds on competing with the oracle rate **11**

A Proofs of information theoretic facts **16**

B Proof of Theorem 16 **17**

C Proofs from Section 4 **19**
- C.1 Proof of Theorem 24 . . . 20
- C.2 Proof of Theorem 25 . . . 22

D Upper bound in terms of (dual) total correlation: proof of Theorem 9 **24**

E Logarithmic overhead is necessary **28**

F Recovering existing bounds **30**
- F.1 Recovering the bound of Li and Cai (2025) . . . 30
- F.2 Recovering the bound of Austin (Austin (2020)) . . . 31

G Decoupling estimation error and sampling error **32**

Appendix A. Proofs of information theoretic facts

This section completes the proofs of the technical lemmas in Sections 1 and 2. We first prove Lemma 8.

Proof [Proof of Lemma 8] First, observe that

$$\sum_{i=1}^{n} Z_i = nH_1 - H_n = \sum_{i=1}^{n} H(X_i) - H(X_1, \ldots, X_n) = \mathrm{TC},$$

which proves item 1. Lastly, using the chain rule of conditional entropy, we observe that

$$\begin{aligned}\mathrm{DTC}+\mathrm{TC} &= \sum_{i=1}^{n}(H(X_i) - H(X_i \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)) \\ &= \sum_{i=1}^{n}(H(X_i) + H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) - H(X_1, \ldots, X_n)) \\ &= n(H_1 + H_{n-1} - H_n) \\ &= nZ_n.\end{aligned}$$

Combining this equation with item 1, it follows that

$$\mathrm{DTC} = nZ_n - TC = nZ_n - \sum_{i=1}^{n} Z_i,$$

which proves item 2. ■

We now prove Lemma 13.

Proof [Proof of Lemma 13] Recall that $I(X_i; X_{S\setminus\{i\}}) = H(X_i) + H(X_{S\setminus\{i\}}) - H(X_S)$; taking expectations, we find that

$$\begin{aligned}Z_i &= \mathbb{E}_{|S|=i-1, j\notin S}\left[I(X_j; X_S)\right] \\ &= \mathbb{E}_{|T|=i, j\in T}\left[H(X_j) + H(X_{T\setminus\{j\}}) - H(X_S)\right] \\ &= H_1 + H_{i-1} - H_i,\end{aligned}$$

as desired. ■

Appendix B. Proof of Theorem 16

Here we prove our main result showing that the optimal error achievable by any mask schedule is given by the optimal left Riemann approximation of the information curve. This is restated below for convenience:

Theorem 16 *Let μ denote the underlying distribution of data. Let $(k, \{s_i\}_1^k)$ be an unmasking schedule and $\{S_i\}_1^k, |S_i| = s_i$ be a fixed subset schedule. Let $N_i = \sum_{j=1}^{i} s_j$ denote the partial sums of the s_i sequence, where $N_0 = 0$. Suppose the fixed unmasking algorithm $\mathcal{A}_{\text{fixed}}(\{S_i\}_1^k)$ samples a distribution $\nu^{S_1,\ldots,S_k}$ and the random unmasking algorithm $\mathcal{A}(k, \{s_i\}_1^k)$ samples a distribution ν. Then both algorithms have query complexity k, and ν achieves KL error relative to μ of*

$$\mathsf{KL}(\mu \,\|\, \nu) \le \mathbb{E}_{S_1,\ldots,S_k}\left[\mathsf{KL}\big(\mu \,\big\|\, \nu^{S_1,\ldots,S_k}\big)\right] = \sum_{i=1}^{k}\left(\sum_{j=1}^{s_i}\left(Z_{N_{i-1}+j} - Z_{N_{i-1}+1}\right)\right),$$

where the expectation is taken over all partitions $S = \bigsqcup_{i=1}^{k} S_i$ for which $|S_i| = s_i$.

Proof Let $T_i = \bigcup_{j<i} S_i$ with $T_1 = \emptyset$ be the coordinates which have already been sampled at stage i. We first work with the distribution $\nu^{S_1,\dots,S_k}$. Observe that

$$\begin{aligned}
\mathsf{KL}\big(\mu \,\big\|\, \nu^{S_1,\dots,S_k}\big) &= \mathbb{E}_{\mathbf{x}\sim\mu}\left[\log \frac{f_\mu(\mathbf{x})}{f_\mu^{\otimes}(\mathbf{x})}\right] \\
&= \mathbb{E}_{\mathbf{x}\sim\mu}\left[\sum_{i=1}^{k} \log \frac{f_{\mu(X_{S_i}|X_{T_i}=x_{T_i})}}{f_{\mu^{\otimes}(X_{S_i}|X_{T_i}=x_{T_i})}}\right] \\
&= \sum_{i=1}^{k} \mathbb{E}_{\mathbf{x}\sim\mu}\left[\mathsf{KL}\big(\mu(X_{S_i} \mid X_{T_i}=x_{T_i}) \,\big\|\, \mu^{\otimes}(X_{S_i} \mid X_{T_i}=x_{T_i})\big)\right].
\end{aligned}$$

To simplify this, observe that the inner KL term is essentially a total correlation of the conditional distribution of X_{S_i} given $X_{T_i} = x_{T_i}$. Therefore, it follows that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mu}\left[\mathsf{KL}\big(\mu(X_{S_i} \mid X_{T_i}=x_{T_i}) \,\big\|\, \mu^{\otimes}(X_{S_i} \mid X_{T_i}=x_{T_i})\big)\right] \\
&\quad= \mathbb{E}_{\mathbf{x}\sim\mu}\left[\left(\sum_{j\in S_i} H(X_j \mid X_{T_i}=x_{T_i})\right) - H(X_{S_i} \mid X_{T_i}=x_{T_i})\right] \\
&\quad= \left(\sum_{j\in S_i} H(X_j \mid X_{T_i})\right) - H(X_{S_i} \mid X_{T_i}) \\
&\quad= \left(\sum_{j\in S_i} H(X_{T_i\cup\{j\}}) - H(X_{T_i})\right) - \left(H(X_{S_i\sqcup T_i}) - H(X_{T_i})\right),
\end{aligned}$$

where in the second equality $H(X_j \mid X_{T_i} = x_{T_i})$ denotes the entropy of the conditional distribution of X_j given $X_{T_i} = x_{T_i}$, while $H(X_j \mid X_{T_i})$ denotes the conditional entropy of X_j given X_{T_i}.

Combining this with the previous equation, we find that

$$\mathsf{KL}\big(\mu \,\big\|\, \nu^{S_1,\dots,S_k}\big) = \sum_{i=1}^{k}\left[\left(\sum_{j\in S_i} H(X_{T_i\cup\{j\}}) - H(X_{T_i})\right) - \left(H(X_{S_i\sqcup T_i}) - H(X_{T_i})\right)\right].$$

Recall now that ν is given by the mixture

$$\nu = \frac{1}{\binom{n}{s_1\dots s_k}} \sum_{\{S_i\}_1^k, |S_i|=s_i} \nu^{S_1,\dots,S_k}.$$

We therefore find that

$$\begin{aligned}
\mathsf{KL}(\mu \,\|\, \nu) &= \mathsf{KL}\left(\mu \,\middle\|\, \frac{1}{\binom{n}{s_1\dots s_k}} \sum_{\{S_i\}_1^k, |S_i|=s_i} \nu^{S_1,\dots,S_k}\right) \\
&\le \frac{1}{\binom{n}{s_1\dots s_k}} \sum_{\{S_i\}_1^k, |S_i|=s_i} \mathsf{KL}\big(\mu \,\big\|\, \nu^{S_1,\dots,S_k}\big) \\
&= \mathbb{E}_{\{S_i\}_1^k, |S_i|=s_i}\left[\mathsf{KL}\big(\mu \,\big\|\, \nu^{S_1,\dots,S_k}\big)\right] \\
&= \mathbb{E}_{\{S_i\}_1^k, |S_i|=s_i}\left[\sum_{i=1}^{k}\left[\left(\sum_{j\in S_i} H(X_{T_i\cup\{j\}}) - H(X_{T_i})\right) - \left(H(X_{S_i\sqcup T_i}) - H(X_{T_i})\right)\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{k} \left[s_i \left(\mathbb{E}_{|S|=N_{i-1}+1} \left[H(X_S) \right] - \mathbb{E}_{|S|=N_{i-1}} \left[H(X_S) \right] \right) - \left(\mathbb{E}_{|S|=N_i} \left[H(X_S) \right] - \mathbb{E}_{|S|=N_{i-1}} \left[H(X_S) \right] \right) \right] \\
&= \sum_{i=1}^{k} \left[s_i \left(H_{N_{i-1}+1} - H_{N_{i-1}} \right) - \left(H_{N_i} - H_{N_{i-1}} \right) \right] \\
&= \sum_{i=1}^{k} \left[s_i H_1 - s_i Z_{N_{i-1}+1} - \sum_{j=1}^{s_i} (H_{N_{i-1}+j} - H_{N_{i-1}+j-1}) \right] \\
&= \sum_{i=1}^{k} \left[\left(\sum_{j=1}^{s_i} Z_{N_{i-1}+j} \right) - s_i Z_{N_{i-1}+1} \right],
\end{aligned}
$$

where the first line is an equality, the second by convexity of KL, the third and fourth lines are direct simplification, the fifth line follows from the fact that $T_i \cup \{j\}$, T_i, S_i, and $S_i \sqcup T_i$ are individually uniformly random subsets of $[n]$ of sizes $N_{i-1} + 1$, N_{i-1}, s_i, and N_i, respectively, and the final three lines are via directly applying the definitions of H_i and $\mathbf{Z}$. The theorem follows from the first, third, and final lines. ■

Now we provide the proof of Theorem 4, restated here for convenience:

Theorem 4 *[Optimal schedule given by best step approximation] Let μ be any distribution over Σ^n, and let $1 \leqslant k \leqslant n$. Let $\mathbf{N}^{*,k}$ be the solution to Eq. (1) for μ's information curve $\mathbf{Z} = \mathbf{Z}(\mu)$.*

Then for any unmasking schedule $s_1, \ldots, s_k$, the expected KL error is given by

$$
\mathop{\mathbb{E}}_{S_1,\ldots,S_k} [\mathsf{KL}(\mu \,\|\, \nu^{S_1,\ldots,S_k})] = \|\mathbf{Z} - \mathbf{Z}^{\mathbf{N}}\|_{L^1}, \qquad \textit{for} \quad N_a \coloneqq 1 + \sum_{t=1}^{a-1} s_t \quad \forall\, a \in [k]\,.
$$

In particular, the schedule that minimizes the expected KL error is

$$
s_t = N^{*,k}_{t+1} - N^{*,k}_{t}, \qquad t \in [k]\,.
$$

Proof Let $\mathbf{N}_a = 1 + \sum_{t=1}^{a-1} s_t \;\; \forall a \in [k]$ and $\mathbf{N}_0 = 1$. By Theorem 16, and the definition of $\mathbf{Z}^{\mathbf{N}}$, we have that

$$
\mathbb{E}_{S_1,\ldots,S_k} \left[\mathsf{KL}\left(\mu \,\middle\|\, \nu^{S_1,\ldots,S^k} \right) \right] = \sum_{i=1}^{k} \left(\sum_{j=1}^{s_i} (Z_{N_{i-1}+j} - Z_{N_{i-1}+1}) \right) = \|\mathbf{Z} - \mathbf{Z}^{\mathbf{N}}\|_{L^1},
$$

yielding the formula for KL error. The remainder of the theorem statement is obvious. ■

Appendix C. Proofs from Section 4

In this section we prove the query complexity lower bounds from Section 4. For the analysis, it will be convenient to use the following interpretation of sampling algorithms in terms of *stochastic decision trees*:

Definition 26 (Stochastic decision tree representation) *Any sampling algorithm $\mathcal{A}$ can be regarded as an (infinite-degree) stochastic decision tree as follows. Every internal node is either a* decision node *(including the root), a* query node*, or a* leaf node*. Decision and leaf nodes (resp. query nodes) are at even (resp. odd) distance from the root:*

- *For every decision node v, the outgoing edges (v, w) connect v to query nodes w. Each such edge is labeled with a partial assignment $X_{S^{(w)}} = x^{(w)}$ with which to query the oracle. From v, the sampler transitions to w with some probability $\mathbb{P}_{\mathcal{A}}[w \mid v]$.*
- *For every query node w, there is a continuum of infinitely many outgoing edges (w, v'), each labeled by an element of $\Delta(\Sigma)^{n-|S^{(w)}|}$ corresponding to a possible response by the oracle to the query $X_{S^{(w)}} = x^{(w)}$. From w, the sampler walks along the edge corresponding to the oracle's response to $X_{S^{(w)}} = x^{(w)}$.*
- *Each leaf node ℓ is labeled with a distribution ν_ℓ over Σ^n, corresponding to the algorithm's (randomized) output if it has reached that state and decided to exit out of the loop. Let $\mathsf{leaf}(\mu)$ (resp. $\mathsf{leaf}^{\leqslant T}(\mu)$) denote all possible leaf nodes of the stochastic decision tree corresponding to $\mathcal{A}$ (resp. which are distance at most $2T$ from the root and reachable given oracle access to μ).*

Every path from the root to a decision or leaf node v is given by a path whose edges are alternatingly labeled by partial assignments $X_S = x$ and corresponding oracle responses.

For any internal or leaf node v of the tree, let $\mathbb{P}_{\mathcal{A}}[v \mid \mu]$ denote the probability that the algorithm traverses that node at some point in its execution, conditioned on the oracle responses coming from the conditional marginal oracle for μ.

C.1. Proof of Theorem 24

We restate the result here for convenience:

Theorem 24 *No $\mathcal{F}$-aware sampling algorithm $\mathcal{A}$ can achieve $\sup_{\mu\in\mathcal{F}} \mathsf{cost}^{\mathsf{TV}}_{\mathcal{T}}(\mathcal{A};\mu) \leqslant 1/16$ for any budget $\mathcal{T}$ satisfying $\mathcal{T}(\mu) \lesssim \max(1, \min(\mathrm{TC}(\mu), \mathrm{DTC}(\mu)))\log(n)$ for all $\mu \in \mathcal{F}$.*

We will use the following terminology: in the stochastic decision tree associated to $\mathcal{A}$, a leaf ℓ is said to *miss* a subspace $\mathcal{V}$ if, for all partial assignments labeling edges from the root-to-leaf path to ℓ, either the assignment is of size less than $\dim \mathcal{V}$, or if otherwise there does not exist $x^* \in \mathcal{V}$ consistent with that assignment. Otherwise, ℓ is said to *hit* $\mathcal{V}$.

Proof Let $\mathcal{D}$ denote the mixture distribution over $\mathcal{F}$ given by

$$\frac{1}{2}\delta_{\mathcal{U}} + \frac{1}{2n-2}\sum_{k=1}^{n-1}\mathcal{D}_k$$

where $\mathcal{D}_k$ are as defined in Lemma 21. We will prove the stronger statement that no $\mathcal{F}$-aware sampling algorithm $\mathcal{A}$ can even achieve $\mathbb{E}_{\mu\sim\mathcal{D}}[\mathsf{cost}^{\mathsf{TV}}_{\mathcal{T}}(\mathcal{A};\mu)] \le 1/4$.

In order for $\mathsf{cost}^{\mathsf{TV}}_{\mathcal{T}}(\mathcal{A};\mu)$ to be finite, we must have $\mathsf{leaf}(\mu) = \mathsf{leaf}^{\leqslant \mathcal{T}(\mu)}(\mu)$. Henceforth, let $\mathsf{leaf}^* := \mathsf{leaf}^{\leqslant \mathcal{T}(\mathcal{U})}(\mathcal{U})$. We must have

$$\mathsf{TV}\Big(\mathcal{U}, \sum_{\ell\in\mathsf{leaf}^*} \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}]\cdot \nu_\ell\Big) \leqslant 1/8\,,$$

or else $\mathbb{E}_{\mu\sim\mathcal{D}}[\mathsf{cost}^{\mathsf{TV}}_{\mathcal{T}}(\mathcal{A};\mu)] \ge \frac{1}{2}\mathsf{cost}^{\mathsf{TV}}_{\mathcal{T}}(\mathcal{A};\mathcal{U}) > 1/16$.

For any leaf node ℓ, let $v_1 \to w_1 \to v_2 \to \cdots \to w_{T-1} \to v_T$ denote the sequence of decision and leaf nodes along the root-to-leaf path to ℓ, and suppose the edges (v_i, w_i) are labeled with partial assignments $X_{S^{(i)}} = x^{(i)}$. If $\ell \in \mathsf{leaf}^*$, then the edges (w_i, v_{i+1}) are all labeled with $\mathrm{Unif}(\mathbb{F}_q)^{\otimes(n-|S^{(i)}|)}$.

Let $k_1 \leqslant \cdots \leqslant k_T$ denote the numbers $|S^{(1)}|, \ldots, |S^{(T)}|$ in sorted order. By Proposition 20, for any MDS $\mathcal{V}$ of dimension $k > k_T$ we have $\mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}_{\mathcal{V}}] = \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}]$. For $k_j < k < k_{j+1}$, by Lemma 21,

$$\mathbb{P}_{\mathcal{V}\sim\mathcal{D}_k}[\ell \text{ avoids } \mathcal{V}] \ge 1 - \sum_{s>j} q^{-(k_s-k)} \ge 1 - T/q\,, \tag{2}$$

and if ℓ avoids $\mathcal{V}$, the oracle's output under every query along the path is uniform marginals and again we have $\mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}_{\mathcal{V}}] = \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}]$. The same reasoning applies to $k < k_1$.

Let us write

$$\mathbb{E}_{\mathcal{V}\sim\mathcal{D}_k}\mathsf{TV}\Big(\mathcal{U}_{\mathcal{V}}, \sum_{\ell\in\mathsf{leaf}(\mathcal{U}_{\mathcal{V}})} \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}_{\mathcal{V}}]\cdot\nu_\ell\Big) \geq 1/2 - \mathbb{E}_{\mathcal{V}\sim\mathcal{D}_k}\mathsf{TV}\Big(\mathcal{U}, \sum_{\ell\in\mathsf{leaf}(\mathcal{U}_{\mathcal{V}})} \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}_{\mathcal{V}}]\cdot\nu_\ell\Big)$$

where we used that $\mathsf{TV}(\mathcal{U},\mathcal{U}_{\mathcal{V}}) \geq 1/2$ for any proper subspace $\mathcal{V}$. We can rewrite the mixture on the right-hand side as

$$\begin{aligned}\sum_{\ell\in\mathsf{leaf}^*:\text{avoids }\mathcal{V}} &\mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}_{\mathcal{V}}]\cdot\nu_\ell + \sum_{\ell\in\mathsf{leaf}^*:\text{hits }\mathcal{V}} \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}_{\mathcal{V}}]\cdot\nu_\ell + \sum_{\ell\in\mathsf{leaf}(\mathcal{U}_{\mathcal{V}})\backslash\mathsf{leaf}^*} \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}_{\mathcal{V}}]\cdot\nu_\ell \\ &= \sum_{\ell\in\mathsf{leaf}^*} \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}]\cdot\nu_\ell - \sum_{\ell\in\mathsf{leaf}^*:\text{hits }\mathcal{V}} \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}]\cdot\nu_\ell + \sum_{\ell\in\mathsf{leaf}(\mathcal{U}_{\mathcal{V}})\backslash\mathsf{leaf}^*} \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}_{\mathcal{V}}]\cdot\nu_\ell\,, \end{aligned} \tag{3}$$

where we used that for $\ell \in \mathsf{leaf}^*$ that avoid $\mathcal{V}$, $\mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}] = \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}_{\mathcal{V}}]$, and for $\ell \in \mathsf{leaf}^*$ that hit $\mathcal{V}$, it must be that $\mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}_{\mathcal{V}}] = 0$ as the sampler under $\mathcal{U}_{\mathcal{V}}$ must deviate from the path that leads to ℓ. As $\sum_{\ell\in\mathsf{leaf}(\mathcal{U}_{\mathcal{V}})\backslash\mathsf{leaf}^*}\mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}_{\mathcal{V}}] = \sum_{\ell\in\mathsf{leaf}^*:\text{hits}\mathcal{V}}\mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}]$, the TV between $\mathcal{U}$ and the mixture in Eq. (3) is thus upper bounded by $1/8 + \sum_{\ell\in\mathsf{leaf}^*:\text{hits }\mathcal{V}}\mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}]$, and thus

$$\mathbb{E}_{\mathcal{V}\sim\mathcal{D}_k}\mathsf{TV}\Big(\mathcal{U}_{\mathcal{V}}, \sum_{\ell\in\mathsf{leaf}(\mathcal{U}_{\mathcal{V}})} \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}_{\mathcal{V}}]\cdot\nu_\ell\Big) \geq \frac{3}{8} - \sum_{\ell\in\mathsf{leaf}^*:\text{hits }\mathcal{V}} \mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}]\,.$$

We say that $\mathcal{V}$ is *η-good* if it satisfies $\sum_{\ell\in\mathsf{leaf}^*:\text{hits}\mathcal{V}}\mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}] \leqslant \eta$ for some $\eta > 0$. Observe that

$$\begin{aligned}\frac{1}{n-1}\sum_{k=1}^{n-1}\Big\{\sum_{\ell\in\mathsf{leaf}^*}\mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}]\cdot\mathbb{P}_{\mathcal{V}\sim\mathcal{D}_k}[\ell \text{ hits } \mathcal{V}]\Big\} &= \sum_{\ell\in\mathsf{leaf}^*}\mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}]\cdot\frac{1}{n-1}\sum_{k=1}^{n-1}\mathbb{P}_{\mathcal{V}\sim\mathcal{D}_k}[\ell \text{ hits } \mathcal{V}] \\ &\leqslant \sum_{\ell\in\mathsf{leaf}^*}\mathbb{P}_{\mathcal{A}}[\ell \mid \mathcal{U}]\cdot\frac{\mathcal{T}(\mathcal{U}) + (n-1-\mathcal{T}(\mathcal{U}))\mathcal{T}(\mathcal{U})/q}{n-1} \\ &= \frac{\mathcal{T}(\mathcal{U}) + (n-1-\mathcal{T}(\mathcal{U}))\mathcal{T}(\mathcal{U})/q}{n-1} \\ &\leqslant \frac{2\mathcal{T}(\mathcal{U})}{n-1}\end{aligned}$$

where in the second step we used that for any leaf ℓ at distance $2T$ from the root, there are at most T dimensions $0 < k < n$ that are equal to the size of some partial assignment along the root-to-left path to ℓ, and for all other dimensions k, $\mathbb{P}_{\mathcal{V}\sim\mathcal{D}_k}[\ell \text{ hits } \mathcal{V}] \leqslant T/q$ by Eq. (2). By Markov's inequality, we conclude that for $\eta := \frac{4\mathcal{T}(\mathcal{U})}{n-1} \ll 1$,

$$\mathbb{P}_{0<k<n,\mathcal{V}\sim\mathcal{D}_k}[\mathcal{V} \text{ is } \eta\text{-good}] \geq 1/2\,.$$

We conclude that

$$\mathbb{E}_{\mu\sim\mathcal{D}}[\mathsf{cost}^{\mathsf{TV}}_{\mathcal{T}}(\mathcal{A};\mu)] \geq \frac{1}{2}\cdot\mathbb{P}_{0<k<n,\mathcal{V}\sim\mathcal{D}_k}[\mathcal{V} \text{ is } \eta\text{-good}]\cdot\Big(\frac{3}{8} - \eta\Big) \geq \frac{1}{16}$$

as claimed. ■

In fact one sees from the definition of η in the proof above that we have shown the even stronger statement that it is necessary to set the budget $\mathcal{T}(\mathcal{U})$ for the uniform distribution to be *linear* in n for the costs to be sufficiently bounded across all $\mu \in \mathcal{F}$. Intuitively, this comes from the fact that one has to make $\Omega(n)$ queries before one can decisively rule out that μ is supported on a subspace.

C.2. Proof of Theorem 25

We restate the result here for convenience:

Theorem 25 *Let* $\mathbf{Z}$ *be the information curve associated to some distribution with total correlation* TC *and dual total correlation* DTC*. Suppose that* $\min(\mathrm{TC}, \mathrm{DTC}) \log n \ll n$*. Given* $1 \leqslant k < n$*, let* $\mathbf{Z}^{\uparrow k}$ *denote the information curve given by*

$$Z_j^{\uparrow k} := \begin{cases} Z_j & \text{if } j \leqslant k \\ Z_j + \log_2(q) & \text{if } j > k \end{cases}$$

There exist a family $\mathcal{F}$ *of distributions whose information curves are from among* $\{\mathbf{Z}, \mathbf{Z}^{\uparrow 1}, \ldots, \mathbf{Z}^{\uparrow n-1}\}$ *such that for every* k *there is at least one such distribution in* $\mathcal{F}$*, and furthermore for any budget* $\mathcal{T}$ *satisfying* $\mathcal{T}(\mu) \lesssim \max(1, \min(\mathrm{TC}(\mu), \mathrm{DTC}(\mu))) \log(n)$ *for all* $\mu \in \mathcal{F}$*, no* $\mathcal{F}$*-aware sampling algorithm* $\mathcal{A}$ *can achieve* $\sup_{\mu \in \mathcal{F}} \mathsf{cost}_{\mathcal{T}}^{\mathsf{TV}}(\mathcal{A}; \mu) \leqslant 1/16$.

Proof Let μ^* denote a distribution over Σ^n with information curve $\mathbf{Z}$, and let $\mathcal{U}$ and $\mathcal{U}_{\mathcal{V}}$ denote the uniform distribution over $\mathbb{F}_q^n$ and the uniform distribution over subspace $\mathcal{V} \subset \mathbb{F}_q^n$ as before, where $\mathcal{V}$ will range over RS codes. Define $\mu^*[\mathcal{U}] := \mu^* \times \mathcal{U}$ and $\mu^*[\mathcal{U}_{\mathcal{V}}] := \mu^* \times \mathcal{U}_{\mathcal{V}}$, regarded as distributions over $(\Sigma \times \mathbb{F}_q)^n$. If $\mathcal{V}$ has dimension k, then by the linearity of the information curve in the average entropy curve, and by additivity of entropy,

$$\mathbf{Z}(\mu^*[\mathcal{U}]) = \mathbf{Z}(\mu^*) + \mathbf{Z}(\mathcal{U}) = \mathbf{Z}$$

$$\mathbf{Z}(\mu^*[\mathcal{U}_{\mathcal{V}}]) = \mathbf{Z}(\mu^*) + \mathbf{Z}(\mathcal{U}_{\mathcal{V}}) = \mathbf{Z}(\mu^*) + (\mathbb{I}[j > k])_j = \mathbf{Z}^{\uparrow k}.$$

So to construct $\mathcal{F}$, we include $\mu^*[\mathcal{U}]$, and then for every dimension $1 \leqslant k < n$ and every k-dimensional RS code $\mathcal{V}$, we include $\mu^*[\mathcal{U}_{\mathcal{V}}]$, thus satisfying the first condition in the Theorem.

The rest of the proof is nearly identical to that of Theorem 24. It remains to verify that the family $\mathcal{F}$ constructed in the proof of Theorem 25 satisfies the second part of Theorem 25. For this, we closely follow the proof of Theorem 24. By Proposition 20, if one queries the conditional marginal oracle for $\mu^*[\mathcal{U}_{\mathcal{V}}]$ on a partial assignment of size $< k$, then the response will be identical to the one for $\mu^*[\mathcal{U}]$, and likewise if the assignment is of size $> k$ and its projection to the $\mathbb{F}_q^n$ component is incompatible with any element of $\mathcal{V}$. We will use the same *hit* and *miss* terminology from before.

Let $\mathcal{D}_k^{\mu^*}$ denote the distribution over $\mu^*[\mathcal{U}_{\mathcal{V}}]$ where $\mathcal{V}$ is a random k-dimensional RS code. Let $\mathcal{D}$ denote the mixture distribution over $\mathcal{F}$ given by

$$\frac{1}{2}\delta_{\mu^*[\mathcal{U}]} + \frac{1}{2n-2}\sum_{k=1}^{n-1} \mathcal{D}_k^{\mu}.$$

As before, let $\mathsf{leaf}^* := \mathsf{leaf}^{\mathcal{T}(\mu^*[\mathcal{U}])}(\mu^*[\mathcal{U}])$. We must have

$$\mathsf{TV}\Big(\mu^*[\mathcal{U}], \sum_{\ell \in \mathsf{leaf}^*} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]] \cdot \nu_\ell\Big) \leqslant 1/8,$$

or else $\mathbb{E}_{\mu^* \sim \mathcal{D}}[\mathsf{cost}_{\mathcal{T}}^{\mathsf{TV}}(\mathcal{A}; \mu)] > 1/16$.

For any leaf node ℓ, let $v_1 \to w_1 \to v_2 \to \cdots \to w_{T-1} \to v_T$ denote the sequence of decision and leaf nodes along the root-to-leaf path to ℓ, and suppose the edges (v_i, w_i) are labeled with partial assignments $X_{S^{(i)}} = x^{(i)}$. If $\ell \in \mathsf{leaf}^*$, then the edges (w_i, v_{i+1}) are labeled with $\nu \otimes \mathrm{Unif}(\mathbb{F}_q)^{\otimes(n-|S^{(i)}|)}$ for some distribution ν over Σ^n.

Let $k_1 \leqslant \cdots \leqslant k_T$ denote the numbers $|S^{(1)}|, \ldots, |S^{(T)}|$ in sorted order. For any $\mathcal{V}$ of dimension $k > k_T$ we have $\mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}_{\mathcal{V}}]] = \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]]$. For $k_j < k < k_{j+1}$, by Lemma 21, Eq. (2) holds as before, and if

ℓ avoids $\mathcal{V}$, the oracles output under every query along the path is of the form $\nu \otimes \mathrm{Unif}(\mathbb{F}_q)^{\otimes(n-|S|)}$ for some distribution ν. In this case, again we have $\mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}_\mathcal{V}]] = \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]]$. The same reasoning applies to $k < k_1$.

Let us write

$$\mathbb{E}_{\mathcal{V}\sim\mathcal{D}_k} \mathsf{TV}\Big(\mu^*[\mathcal{U}_\mathcal{V}], \sum_{\ell\in\mathsf{leaf}(\mu^*[\mathcal{U}_\mathcal{V}])} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}_\mathcal{V}]]\cdot\nu_\ell\Big) \geq 1/2 - \mathbb{E}_{\mathcal{V}\sim\mathcal{D}_k} \mathsf{TV}\Big(\mu^*[\mathcal{U}], \sum_{\ell\in\mathsf{leaf}(\mu^*[\mathcal{U}_\mathcal{V}])} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}_\mathcal{V}]]\cdot\nu_\ell\Big)$$

where we used that $\mathsf{TV}(\mu^*[\mathcal{U}], \mu^*[\mathcal{U}_\mathcal{V}]) \geq \mathsf{TV}(\mathcal{U}, \mathcal{U}_\mathcal{V}) \geq 1/2$ for any proper subspace $\mathcal{V}$. We can rewrite the mixture on the right-hand side as

$$\begin{aligned}
&\sum_{\ell\in\mathsf{leaf}^*:\text{avoids } \mathcal{V}} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}_\mathcal{V}]]\cdot\nu_\ell + \sum_{\ell\in\mathsf{leaf}^*:\text{hits } \mathcal{V}} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}_\mathcal{V}]]\cdot\nu_\ell + \sum_{\ell\in\mathsf{leaf}(\mu^*[\mathcal{U}_\mathcal{V}])\setminus\mathsf{leaf}^*} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}_\mathcal{V}]]\cdot\nu_\ell \\
&\quad = \sum_{\ell\in\mathsf{leaf}^*} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]]\cdot\nu_\ell - \sum_{\ell\in\mathsf{leaf}^*:\text{hits } \mathcal{V}} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]]\cdot\nu_\ell + \sum_{\ell\in\mathsf{leaf}(\mu^*[\mathcal{U}_\mathcal{V}])\setminus\mathsf{leaf}^*} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}_\mathcal{V}]]\cdot\nu_\ell\,, \quad (4)
\end{aligned}$$

where we used that for $\ell \in \mathsf{leaf}^*$ that avoid $\mathcal{V}$, $\mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]] = \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}_\mathcal{V}]]$, and for $\ell \in \mathsf{leaf}^*$ that hit $\mathcal{V}$, it must be that $\mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}_\mathcal{V}]] = 0$ as the sampler under $\mu^*[\mathcal{U}_\mathcal{V}]$ must deviate from the path that leads to ℓ. As $\sum_{\ell\in\mathsf{leaf}(\mu^*[\mathcal{U}_\mathcal{V}])\setminus\mathsf{leaf}^*} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}_\mathcal{V}]] = \sum_{\ell\in\mathsf{leaf}^*:\text{hits}\mathcal{V}} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]]$, the TV between $\mu^*[\mathcal{U}]$ and the mixture in Eq. (4) is thus upper bounded by $1/8 + \sum_{\ell\in\mathsf{leaf}^*:\text{hits } \mathcal{V}} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]]$, and thus

$$\mathbb{E}_{\mathcal{V}\sim\mathcal{D}_k} \mathsf{TV}\Big(\mu^*[\mathcal{U}_\mathcal{V}], \sum_{\ell\in\mathsf{leaf}(\mu^*[\mathcal{U}_\mathcal{V}])} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}_\mathcal{V}]]\cdot\nu_\ell\Big) \geq \frac{3}{8} - \sum_{\ell\in\mathsf{leaf}^*:\text{hits } \mathcal{V}} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]]\,.$$

We say that $\mathcal{V}$ is η-*good* if it satisfies $\sum_{\ell\in\mathsf{leaf}^*:\text{hits}\mathcal{V}} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]] \leqslant \eta$ for some $\eta > 0$. Observe that

$$\begin{aligned}
&\frac{1}{n-1}\sum_{k=1}^{n-1}\Big\{\sum_{\ell\in\mathsf{leaf}^*} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]]\cdot \mathbb{P}_{\mathcal{V}\sim\mathcal{D}_k}[\ell \text{ hits } \mathcal{V}]\Big\} \\
&\quad = \sum_{\ell\in\mathsf{leaf}^*} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]]\cdot \frac{1}{n-1}\sum_{k=1}^{n-1} \mathbb{P}_{\mathcal{V}\sim\mathcal{D}_k}[\ell \text{ hits } \mathcal{V}] \\
&\quad \leqslant \sum_{\ell\in\mathsf{leaf}^*} \mathbb{P}_{\mathcal{A}}[\ell \mid \mu^*[\mathcal{U}]]\cdot \frac{\mathcal{T}(\mu^*[\mathcal{U}]) + (n-1-\mathcal{T}(\mu^*[\mathcal{U}]))\mathcal{T}(\mu^*[\mathcal{U}])/q}{n-1} \\
&\quad = \frac{\mathcal{T}(\mu^*[\mathcal{U}]) + (n-1-\mathcal{T}(\mu^*[\mathcal{U}]))\mathcal{T}(\mu^*[\mathcal{U}])/q}{n-1} \leqslant \frac{2\mathcal{T}(\mu^*[\mathcal{U}])}{n-1}
\end{aligned}$$

where in the second step we used that for any leaf ℓ at distance $2T$ from the root, there are at most T dimensions $0 < k < n$ that are equal to the size of some partial assignment along the root-to-left path to ℓ, and for all other dimensions k, $\mathbb{P}_{\mathcal{V}\sim\mathcal{D}_k}[\ell \text{ hits } \mathcal{V}] \leqslant T/q$ by Eq. (2). By Markov's inequality, we conclude that for $\eta := \frac{4\mathcal{T}(\mu^*[\mathcal{U}])}{n-1} \ll 1$ (here we used the hypothesis that $\min(\mathrm{TC}, \mathrm{DTC})\log n \ll n$),

$$\mathbb{P}_{0<k<n, \mathcal{V}\sim\mathcal{D}_k}[\mathcal{V} \text{ is } \eta\text{-good}] \geq 1/2\,.$$

We conclude that

$$\mathbb{E}_{\mu\sim\mathcal{D}}[\mathsf{cost}^{\mathsf{TV}}_{T}(\mathcal{A};\mu)] \geq \frac{1}{2}\cdot \mathbb{P}_{0<k<n, \mathcal{V}\sim\mathcal{D}_k}[\mathcal{V} \text{ is } \eta\text{-good}]\cdot\Big(\frac{3}{8} - \eta\Big) \geq \frac{1}{16}\,.$$

■

Appendix D. Upper bound in terms of (dual) total correlation: proof of Theorem 9

In this section, we use Theorem 16 to obtain data-agnostic bounds for the expected KL error of the fixed and random unmasking algorithms. As mentioned in Section 1.4, Theorem 16 already immediately implies bounds from the prior work of Li and Cai (2025) and Austin (2020). In this section we use Theorem 16 to improve these bounds in most regimes, assuming only access to estimates $\widehat{\mathrm{TC}}$ and $\widehat{\mathrm{DTC}}$ of TC and DTC.

Recall that Theorem 9 states the existence of an algorithm attaining error at most ε and query complexity

$$k \le 2 + (1+\log n)\cdot(1+\lceil \widehat{\mathrm{TC}}/\varepsilon\rceil) \qquad (\text{resp. } k \leqslant 2 + (1+\log n)\cdot(1+\lceil \widehat{\mathrm{DTC}}/\varepsilon\rceil)\,.$$

Provided that $\widehat{\mathrm{TC}}$ and $\widehat{\mathrm{DTC}}$ are constant-factor approximations of their respective estimands, this yields a query complexity proportional to $\min(\mathrm{TC}, \mathrm{DTC})$ and is generally significantly better than Theorems 34 and 31.

We now turn to the main technical content of this section, namely proving this result.

Proof [Proof of Theorem 9] We split into two cases, which are roughly similar. The main idea is to use an exponentially increasing schedule to attain the $\widehat{\mathrm{DTC}}$ bound and an exponentially decreasing schedule for the $\widehat{\mathrm{TC}}$ bound. This will attain the correct query complexity. Moreover, using the pictorial representation, we find that the horizontal slices of the error can be enlarged by a factor of $\frac{\widehat{\mathrm{DTC}}}{\varepsilon}$ or $\frac{\widehat{\mathrm{TC}}}{\varepsilon}$, respectively, and subsequently shifted horizontally to fit above or below the information curve, respectively. From this it follows that the total error is at most a factor of $\frac{\varepsilon}{\widehat{\mathrm{DTC}}}$ or $\frac{\varepsilon}{\widehat{\mathrm{TC}}}$ times either the area DTC or the area TC, respectively, yielding the upper bound of ε, provided that $\mathrm{TC} \le \widehat{\mathrm{TC}}$ and $\mathrm{DTC} \le \widehat{\mathrm{DTC}}$. We provide the full details below.

1. The $\widehat{\mathrm{TC}}$ bound. We proceed by defining the mask schedule, and then analyzing the query complexity and sampling error.

Mask Schedule. We first define our mask schedule. Let $\zeta = 1 + \left\lceil \frac{\widehat{\mathrm{TC}}}{\varepsilon} \right\rceil > 1$. If $\zeta \ge n+1$, then pick $k = n$ and $s_i = 1$ for all i; the sampler is perfect and the query complexity is $n \le \left\lceil 1 + \frac{\widehat{\mathrm{TC}}}{\varepsilon} \right\rceil$, resolving this special case. From now on assume $\zeta \le n$. Consider the sequence N_i given by $N_0 = 0$ and then recursively

$$N_i = \left\lfloor N_{i-1} + (n - N_{i-1})\frac{1}{\zeta} \right\rfloor$$

for $1 \le i \le \left\lfloor \frac{\log(n-\zeta+1)}{\log \frac{1}{1-\frac{1}{\zeta}}} \right\rfloor + 2 = \lambda$.

Note that definitionally we have $N_i \ge N_{i-1}$ and by induction $N_i \le n-1$ for all i. Moreover,

$$N_i \ge N_{i-1} + (n - N_{i-1})\frac{1}{\zeta} - \frac{\zeta - 1}{\zeta} = (N_{i-1} - 1)\left(1 - \frac{1}{\zeta}\right) + n\frac{1}{\zeta}\,,$$

so that

$$n - \zeta + 1 - N_i \le (n - \zeta + 1 - N_{i-1})\left(1 - \frac{1}{\zeta}\right).$$

It follows that

$$n - 1 \ge N_\lambda \ge (n-\zeta+1)\left(1 - \left(1 - \frac{1}{\zeta}\right)^{\lambda}\right) > (n - \zeta + 1)\left(1 - \frac{1}{n-\zeta+1}\right) = n - \zeta.$$

Now set $N_i = N_{i-1} + 1$ for $\lambda + 1 \le i \le \lambda + n - N_\lambda$. Note that $N_{\lambda+n-N_\lambda} = n$. Lastly, define $s_i = N_i - N_{i-1}$.[4] We consider the mask schedule given by $\{s_i\}_1^{\lambda+n-N_\lambda}$.

Query Complexity. The query complexity k equals the number of steps of unmasking, i.e.

$$k = \lambda + n - N_\lambda \le \zeta + 2 + \frac{\log n}{\log \frac{1}{1-\frac{1}{\zeta}}} \le 2 + \zeta(1 + \log n) \le 2 + (1 + \log n) \cdot \left[1 + \left\lceil \frac{\widehat{\mathrm{TC}}}{\varepsilon} \right\rceil\right],$$

where we have used the fact that $\log \frac{1}{1-z} = -\log(1-z) \ge z$ for $z = \frac{1}{\zeta} \in [0, 1)$.

Sampling Error. First observe that for $1 \le i \le \lambda$, we have

$$s_i = N_i - N_{i-1} \le (n - N_{i-1})\frac{1}{\zeta} \implies s_i \le (n - N_i)\frac{1}{\zeta - 1} \le \frac{\varepsilon}{\widehat{\mathrm{TC}}}(n - N_i).$$

Applying Theorem 16, we find that

$$\begin{aligned}
\mathsf{KL}(\mu \,\|\, \nu) &\le \mathbb{E}_{S_1,\ldots,S_k}[\mathsf{KL}(\mu \,\|\, \nu^{S_1,\ldots,S_k})] \\
&= \sum_{i=1}^{k} \left[\left(\sum_{j=1}^{s_i} Z_{N_{i-1}+j} \right) - s_i Z_{N_{i-1}+1} \right] \\
&\le \sum_{i=1}^{k} s_i \left(Z_{N_i} - Z_{N_{i-1}+1} \right) \\
&\le \frac{\varepsilon}{\widehat{\mathrm{TC}}} \left(\sum_{i=1}^{\lambda} (n - N_i)(Z_{N_i} - Z_{N_{i-1}}) \right) + \sum_{i=\lambda+1}^{k} s_i (Z_{N_i} - Z_{N_{i-1}+1}) \\
&\le \frac{\varepsilon}{\widehat{\mathrm{TC}}} \left(\sum_{i=1}^{\lambda-1} (N_{i+1} - N_i) Z_{N_i} \right) + \frac{\varepsilon}{\widehat{\mathrm{TC}}}(n - N_\lambda) Z_{N_\lambda} \\
&\le \frac{\varepsilon}{\widehat{\mathrm{TC}}} \left(\sum_{i=1}^{\lambda-1} \sum_{j=N_i}^{N_{i+1}-1} Z_j \right) + \frac{\varepsilon}{\widehat{\mathrm{TC}}} \sum_{j=N_\lambda}^{n-1} Z_j \\
&= \frac{\varepsilon}{\widehat{\mathrm{TC}}} \sum_{j=1}^{n-1} Z_j \\
&\le \frac{\varepsilon}{\widehat{\mathrm{TC}}} \cdot \mathrm{TC} \\
&\le \varepsilon,
\end{aligned}$$

where we let $Z_0 = Z_1 = 0$ by convention and we have repeatedly used that the Z_j's are nonnegative and nondecreasing (see Lemma 8). Note that the fourth line follows from a rearrangement and the fact that $N_i = N_{i-1} + 1$ for $i > \lambda$. Thus the algorithm yields the correct query complexity and sampling error, completing the proof of this case.

2. The $\widehat{\mathrm{DTC}}$ bound. We proceed in the same three steps as in case 1; the proof will be largely similar, except that the mask schedule is essentially flipped.

Mask Schedule. Let $\zeta = 1 + \left\lceil \frac{\widehat{\mathrm{DTC}}}{\varepsilon} \right\rceil > 1$. If $\zeta \ge n + 1$, then pick $k = n$ and $s_i = 1$ for all i; the sampler is perfect and the query complexity is $n \le \lceil 1 + \frac{\mathrm{DTC}}{\varepsilon} \rceil$, resolving this special case. From now on assume

4. Note that potentially some of the final values of s_i will be 0.

$\zeta \leq n$. Consider the sequence N_i' given by $N_0' = n$ and then recursively

$$N_i' = \left\lceil N_{i-1}'\left(1 - \frac{1}{\zeta}\right)\right\rceil$$

for $1 \leq i \leq \left\lfloor \frac{\log(n-\zeta+1)}{\log\frac{1}{1-\frac{1}{\zeta}}} \right\rfloor + 2 = \lambda$.

Note that definitionally we have $N_i' \leq N_{i-1}'$ and by induction $N_i' \geq 1$ for all i. Moreover,

$$N_i' \leq N_{i-1}'\left(1 - \frac{1}{\zeta}\right) + \frac{\zeta - 1}{\zeta} \implies N_i' - \zeta + 1 \leq \left(N_{i-1}' - \zeta + 1\right)\left(1 - \frac{1}{\zeta}\right).$$

It follows that

$$1 \leq N_\lambda' \leq \zeta - 1 + (n - \zeta + 1)\left(1 - \frac{1}{\zeta}\right)^\lambda < \zeta - 1 + (n - \zeta + 1)\frac{1}{n - \zeta + 1} = \zeta.$$

Now, set $N_i' = N_{i-1}' - 1$ for $\lambda + 1 \leq i \leq \lambda + N_\lambda'$. Note that $N_{\lambda+N_\lambda'}' = 0$. Lastly, define $s_i = N_{\lambda+N_\lambda'-i}' - N_{\lambda+N_\lambda'-i+1}'$. We consider the mask schedule given by $\{s_i\}_1^{\lambda+N_\lambda'}$.

Query Complexity. The query complexity k equals the number of steps of unmasking, i.e.

$$k = \lambda + N_\lambda' \leq \zeta + 2 + \frac{\log n}{\log\frac{1}{1-\frac{1}{\zeta}}} \leq 2 + \zeta(1 + \log n) \leq 2 + (1 + \log n) \cdot \left\lfloor 1 + \left\lceil \frac{\widehat{\mathrm{DTC}}}{\varepsilon} \right\rceil \right\rfloor,$$

where we have used the fact that $\log\frac{1}{1-z} = -\log(1 - z) \geq z$ for $z = \frac{1}{\zeta} \in [0, 1)$.

Sampling Error. First observe that for $i > N_\lambda'$, we have $\lambda + N_\lambda' - i + 1 \leq \lambda$ and hence

$$s_i = N_{\lambda+N_\lambda'-i}' - N_{\lambda+N_\lambda'-i+1}' \leq N_{\lambda+N_\lambda'-i}'\frac{1}{\zeta} \implies s_i \leq N_{\lambda+N_\lambda'-i+1}'\frac{1}{\zeta - 1} \leq \frac{\varepsilon}{\widehat{\mathrm{DTC}}}N_{\lambda+N_\lambda'-i+1}'.$$

As usual, let

$$N_i = \sum_{j \leq i} s_i = N_{\lambda+N_\lambda'-i}'.$$

Applying Theorem 16, we find that

$$\begin{aligned}
\mathsf{KL}(\mu \,\|\, \nu) &\leq \mathbb{E}_{S_1,\ldots,S_k}[\mathsf{KL}\big(\mu \,\big\|\, \nu^{S_1,\ldots,S_k}\big)] \\
&= \sum_{i=1}^k \left[\left(\sum_{j=1}^{s_i} Z_{N_{i-1}+j}\right) - s_i Z_{N_{i-1}+1}\right] \\
&\leq \sum_{i=1}^k s_i(Z_{N_i} - Z_{N_{i-1}+1}) \\
&\leq \frac{\varepsilon}{\widehat{\mathrm{DTC}}}\left(\sum_{i=N_\lambda'+1}^{N_\lambda'+\lambda} N_{\lambda+N_\lambda'-i+1}'(Z_{N_i} - Z_{N_{i-1}})\right) + \sum_{i \leq N_\lambda'} s_i(Z_{N_i} - Z_{N_{i-1}+1}) \\
&= \frac{\varepsilon}{\widehat{\mathrm{DTC}}}\left(N_{N_\lambda'+\lambda-1}Z_n\right) - \frac{\varepsilon}{\widehat{\mathrm{DTC}}}\left(\sum_{i=N_{\lambda'+1}}^{N_\lambda'+\lambda-1}(N_i - N_{i-1})Z_{N_i}\right) - \frac{\varepsilon}{\widehat{\mathrm{DTC}}}\left(N_{N_\lambda'}Z_{N_{N_\lambda'}}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\varepsilon}{\widehat{\mathrm{DTC}}}\left(N_{N'_\lambda+\lambda-1}Z_n\right) - \frac{\varepsilon}{\widehat{\mathrm{DTC}}}\left(\sum_{i=N_{\lambda'+1}}^{N'_\lambda+\lambda-1}\left(\sum_{j=N_{i-1}+1}^{N_i} Z_j\right)\right) - \frac{\varepsilon}{\widehat{\mathrm{DTC}}}\left(\sum_{j=1}^{N_{N'_\lambda}} Z_j\right) \\
&= \frac{\varepsilon}{\widehat{\mathrm{DTC}}}\left(\sum_{j=1}^{N_{N'_\lambda+\lambda-1}}(Z_n - Z_j)\right) \\
&\leq \frac{\varepsilon}{\widehat{\mathrm{DTC}}}\sum_{j=1}^{n}(Z_n - Z_j) \\
&= \frac{\varepsilon}{\widehat{\mathrm{DTC}}} \cdot \mathrm{DTC} \\
&\leq \varepsilon,
\end{aligned}
$$

where we let $Z_0 = Z_1 = 0$ by convention and we have repeatedly used that the Z_j's are nonnegative and nondecreasing (see Lemma 8). Note that the fourth line follows from a rearrangement and the fact that $N_i = N_{i-1} + 1$ for $i \leq N'_\lambda$. Thus the algorithm yields the correct query complexity and sampling error, completing the proof of this case.

In conclusion, we find that in both cases there exists a mask schedule with the desired query complexity and sampling error. This completes the proof. ■

Knowledge of TC and DTC. We elaborate briefly on the "hyperparameter sweep" discussed in the introduction. While the above result does not require knowledge of the data distribution or the entire information curve, it nonetheless requires the values of TC and DTC. These values are in general unknown and moreover not readily estimable from our conditional oracle. In practice, however, we can treat TC and DTC as sampling hyperparameters and sweep over a feasible range $\mathcal{H}$.

We suggest a choice of $\mathcal{H}$ as follows. First, it is not difficult to see that if we choose estimates $\widehat{\mathrm{TC}} \in [\mathrm{TC}, 2 \cdot \mathrm{TC}]$ and $\widehat{\mathrm{DTC}} \in [\mathrm{DTC}, 2 \cdot \mathrm{DTC}]$, the mask schedule in the proof of Theorem 9 achieves error at most ε and query complexity within a factor of two of

$$
2 + (1 + \log n) \cdot \left(1 + \min\left(\left\lceil \frac{\mathrm{TC}}{\varepsilon} \right\rceil, \left\lceil \frac{\mathrm{DTC}}{\varepsilon} \right\rceil\right)\right),
$$

that is the complexity if we had complete knowledge of TC and DTC. Moreover, we know that

$$
1 \leq \left\lceil \frac{\mathrm{DTC}}{\varepsilon} \right\rceil, \left\lceil \frac{\mathrm{TC}}{\varepsilon} \right\rceil \quad \text{and} \quad \mathrm{DTC}, \mathrm{TC} \leq nZ_n \leq nH_1 \leq n \log |\Sigma|.
$$

Combining these observations, we can take

$$
\mathcal{H} = \{2^i : i \in \mathbb{Z}, \varepsilon \leq i \leq n \log |\Sigma|\}; \quad |\mathcal{H}| = \mathcal{O}\left(\log \frac{n \log |\Sigma|}{\varepsilon}\right),
$$

for which there exists $(\widehat{\mathrm{TC}}, \widehat{\mathrm{DTC}}) \in \mathcal{H}^2$ which if used as the estimates of TC and DTC yields the desired error and query complexity. Under this choice of $\mathcal{H}$, the hyperparameter sweep incurs an extra query complexity factor of $|\mathcal{H}|^2 = \mathcal{O}\left(\log\left(\frac{n \log |\Sigma|}{\varepsilon}\right)^2\right)$. For most choices of $n, \varepsilon, |\Sigma|$, this will be a polylogarithmic factor in n.

Appendix E. Logarithmic overhead is necessary

In this section, we show that the $\log n$ term in Theorem 9 is necessary. This is essentially implicit in Birgé (1987) and has been used in various works on monotone distribution estimation. We were, however, not able to find a complete statement and proof in the literature and provide one for completeness. First, recall the following definition.

Definition 27 (Piecewise Functions) *A k-piecewise function $f : [n] \to R$ is a function for which there are at most k values $i \in [n-1]$ such that $f(i) \neq f(i+1)$.*

We can now state the main result of this subsection.

Lemma 28 *Let $n \geq 2$, ε be such that $\frac{2}{n}\log\frac{2}{\varepsilon} \leq \varepsilon \leq \frac{1}{\log n}$, and $k \leq c \cdot \frac{\log n}{\varepsilon}$ for some sufficiently small constant c. Then there exists a non-negative monotone increasing function $f : [n] \to \mathbb{R}$ satisfying $\sum_i f(i) = 1$ such that for any function $h : [n] \to \mathbb{R}$ which is a k-piecewise constant function, we have that*

$$\|f - g\|_{L^1} \geq \Omega(\varepsilon).$$

Proof We define f as follows. For $i = 0, \dots, \left\lceil \frac{\log(n+1)}{\log(1+\varepsilon)} - 1 \right\rceil = m \leq \frac{\log(n+1)}{\varepsilon}$, we let $B_i = \{\lfloor (1+\varepsilon)^i \rfloor, \dots, \min(\lfloor (1+\varepsilon)^{i+1} \rfloor - 1, n)\}$, and for $x \in B_i$ we let

$$f(x) = p_i = \frac{1}{4} \cdot \frac{(1+\varepsilon)^{-i}}{\log n}.$$

Note that

$$\sum_{i=1}^{n} f(x) \leq \frac{1}{4} \sum_{i=0, \varepsilon(1+\varepsilon)^i \geq 1}^{m} (\varepsilon(1+\varepsilon)^i + 1) \frac{(1+\varepsilon)^{-i}}{\log n} \leq \frac{1}{4} \left(\sum_{i=0}^{m} 2 \frac{\varepsilon}{\log n} \right) \leq 1$$

as $m \leq \frac{2 \log n}{\varepsilon}$ and so this is a valid choice of f.

Let h be any k-piecewise function, and let $I_1, \dots, I_k$ denote the partition of $[n]$ into k disjoint intervals on which h is constant. We will refer to the right endpoints of these intervals as the *breakpoints* of h. The remainder of the proof proceeds in two main steps. We first show that h may be modified to have two desirable structural qualities: namely to have breakpoints only at the right endpoints of the B_i and values only in the $\{p_i\}$. We then directly analyze functions h with these properties to prove the bound.

Shifting the image of h. We first claim that for any h, there is some h_{image} such that $\|h_{\text{image}} - f\|_{L^1} \leq \|h - f\|_{L^1}$, and h_{image} has the same breakpoints as h but $h_{\text{image}}(x) = p_{i_j}$ for all $x \in I_j$, for values i_j satisfying $B_{i_j} \cap I_j \neq \emptyset$. That is, on each interval where h is constant, we can replace the value of h on that interval with one of the values that f attains on the same interval.

Indeed, this is because the contribution of I_j to the total error is $F_j(h(x)) = \sum_{x \in I_j} |h(x) - f(x)|$. Suppose $h'(x) \in [p_t, p_{t+1}]$. Then since $F_j(h'(x))$ is linear in $[p_t, p_{t+1}]$, it must be optimized at one of the endpoints. In particular, we may replace the value of h on I_j with either p_t or p_{t+1} without increasing the number of pieces of total error. Applying this result to all intervals I_j yields some h_{image} which satisfies the conditions of the claim.

Shifting the breakpoints of h. We now claim that for any h, there is some h_{final} such that $\|h_{\text{final}} - f\|_{L^1} \leq \|h - f\|_{L^1}$, which is also k-piecewise but whose breakpoints are a subset of the breakpoints of f, and whose image is contained in the image of h.

Indeed, suppose h has a breakpoint $t \in I_j \cap B_\ell$ which is not the right endpoint of B_ℓ; call such breakpoints *bad*. The contribution of I_j to the total error is $F_j(h(x)) = \sum_{x \in I_j} |h(x) - f(x)|$. We now have two cases.

Case 1. If $|h(t) - f(t)| > |h(t+1) - f(t+1)|$, let $h'(r) = h(t+1)\ \forall\, r \in I_j \cap B_\ell, r \le t$, and $h'(x) = h(x)$ otherwise. Then the number of pieces and total error of h' are not greater than those of h. Moreover, either the total number of pieces decreases by 1 or h' has one less bad breakpoint than h.[5]

Case 2. If $|h(t) - f(t)| \le |h(t+1) - f(t+1)|$, let $h'(r) = h(t)\ \forall\, r \in I_{j+1} \cap B_\ell, r > t$, and $h'(x) = h(x)$ otherwise. Then the number of pieces and total error of h' are not greater than those of h. Moreover, either the total number of pieces decreases by 1 or h' has one less bad breakpoint than h.[6]

Iterating the operation $h \to h'$, we conclude that after a finite number of rounds there are no bad breakpoints. The output h_{final} satisfies the conditions of the claim.

Completing the proof. Combining these claims, there exists *an* optimal k-piecewise approximation h to f for which $I_j = \{\lfloor (1+\varepsilon)^{i_j} \rfloor + 1, \lfloor (1+\varepsilon)^{i_{j+1}} \rfloor\}$ for some $0 = i_1 \le \cdots \le i_{k+1} = m$, and for all $x \in I_j$, we have that $h(x) = p_{t_j}$ for some t_j. Moreover, breaking up each interval into the intervals before and after t_j, we may further assume that $h(x) = p_{i_j}$ or $h(x) = p_{i_{j+1}}$.[7] Lastly, we may ignore the regions B_j for which $j \lesssim \frac{\log(2/\varepsilon)}{\varepsilon}$, so that for all considered B_s, we have $\varepsilon(1+\varepsilon)^s \ge 2$.[8] Now, let $\ell_j = i_{j+1} - i_j$. We have the following two cases.

Large interval regime. First, if there is any interval I_j such that $\ell_j \ge \frac{2(1+\varepsilon)}{\varepsilon}$, then if $h(x) = (1+\varepsilon)^{i_{j+1}}$ on this interval, and so the error of the approximation on this interval is at least

$$\begin{aligned} \frac{1}{\log n} \sum_{s=i_j}^{i_{j+1}} \left| \frac{1}{(1+\varepsilon)^s} - \frac{1}{(1+\varepsilon)^{i_{j+1}}} \right| \cdot \frac{1}{2}\varepsilon(1+\varepsilon)^s &= \frac{\varepsilon}{2\log n} \left(\ell_j - \sum_{s=0}^{\ell_j - 1} \frac{1}{(1+\varepsilon)^s} \right) \\ &\ge \frac{\varepsilon}{2\log n} \left(\ell_j - \frac{1+\varepsilon}{\varepsilon} \right) \\ &\ge \tfrac{1}{2\log n} = \Omega(\varepsilon). \end{aligned}$$

If instead $h(x) = (1+\varepsilon)^{i_j}$, we obtain the same result via an analogous calculation. In either case, the desired statement holds.

Small interval regime. Otherwise, assume that $\ell_j \le \frac{2(1+\varepsilon)}{\varepsilon}$ for all j. Recall the approximation inequality $1 - \frac{1}{(1+\varepsilon)^r} \ge \min(1/2, r\varepsilon/2)$, which follows from Bernoulli's inequality. Applying it, we find that if $h(x) = (1+\varepsilon)^{i_{j+1}}$, the error on I_j is at least

$$\begin{aligned} \frac{1}{\log n} \sum_{s=i_j}^{i_{j+1}} \left| \frac{1}{(1+\varepsilon)^s} - \frac{1}{(1+\varepsilon)^{i_{j+1}}} \right| \cdot \frac{1}{2}\varepsilon(1+\varepsilon)^s &\ge \frac{\varepsilon}{4\log n} \sum_{s=0}^{\ell_j - 1} s\varepsilon \\ &\ge \frac{\varepsilon^2}{16\log n} (\ell_j - 1)^2 \end{aligned}$$

If instead $h(x) = (1+\varepsilon)^{i_j}$, we obtain the same result again via an analogous calculation. Thus, the overall error of the approximation can be lower bounded by

$$\|h - f\|_{L^1} \ge \frac{\varepsilon^2}{16\log n} \sum_{j=1}^{k} (\ell_j - 1)^2.$$

5. The former occurs if I_j has a smaller left endpoint than B_ℓ and the latter otherwise.
6. The former occurs if I_{j+1} has a larger right endpoint than B_ℓ and the latter otherwise.
7. Note that this operation at most doubles the number of intervals in h; thus, it can be accounted for by changing the constant c appropriately.
8. This operation at most increments the number of intervals by a constant, and underestimates the total error, so it is also permissible.

Since $\sum_{j=1}^{k}(\ell_j - 1) \geq (1-c)\frac{\log n}{\varepsilon}$, by the Cauchy-Schwarz inequality we conclude that

$$\|h - f\|_{L^1} \geq \frac{\varepsilon^2(1-c)^2}{16\log n} \cdot \frac{(\log n)^2}{k\varepsilon^2} = \frac{(1-c)^2}{16} \cdot \frac{\log n}{k} \geq \Omega\left(\varepsilon\right),$$

completing the proof of the desired claim. ■

To make this result applicable to distributions, recall the following result about the realizability of any information curve by some distribution:

Lemma 29 (Theorem 1 in Chen et al. (2009)) *For any information curve $0 \leq Z_1 \leq \cdots \leq Z_n$ and approximation errors $\varepsilon_1, \ldots, \varepsilon_n > 0$, there exists a distribution μ for which $|Z_i(\mu) - Z_i| \leq \varepsilon_i$ for all i.*

Combining Lemmas 28 and 29, we conclude that there exist distributions for which the $\log n$ is unavoidable on many parameter regimes.

Theorem 30 *Let $n \geq 2$, ε be such that $\frac{2}{n}\log\frac{2}{\varepsilon} \leq \varepsilon \leq \frac{1}{\log n}$, and c be a sufficient small constant. Then there exists a distribution μ such that for any unmasking schedule $(k, \{s_i\}_1^k)$ with $k \leq c \cdot \frac{\log n}{\varepsilon}$, we have*

$$\mathbb{E}_{S_1,\ldots,S_k}\left[\mathsf{KL}\left(\mu \,\|\, \nu^{S_1,\ldots,S_k}\right)\right] \geq \Omega(\varepsilon),$$

where the expectation is taken over all partitions $S = \bigsqcup_{i=1}^{k} S_i$ and $\nu^{S_1,\ldots,S_k}$ denotes the distribution outputted by $\mathcal{A}(k, \{S_i\}_1^k)$.

Proof Applying Theorem 4, the expected KL error is given by $\|\mathbf{Z} - \mathbf{Z}^{\mathbf{N}}\|_{L^1}$. The result follows from choosing a distribution μ via Lemma 29 with $\varepsilon_i = o(\frac{\varepsilon}{n})$ and applying Lemma 28 and the triangle inequality for the L^1 norm. ■

Appendix F. Recovering existing bounds

In this section we recover the iteration complexity bound of Li and Cai (2025) and an iteration complexity bound which is implicit in Austin (2020).

F.1. Recovering the bound of Li and Cai (2025)

In Li and Cai (2025), the authors prove the following bound on the sampling error, given the sizes of the mask schedule.

Theorem 31 (Theorem 1 of Li and Cai (2025)) *Let μ be the data distribution and $\nu^{S_1,\ldots,S_k}$ be the output of the fixed unmasking algorithm $\mathcal{A}_{\text{fixed}}(k, \{S_i\}_1^k)$. Let $s_{\max} = \max_{i=1}^{k}|S_i|$. Then we have*

$$\mathbb{E}_{S_1,\ldots,S_k}\left[\mathsf{KL}\left(\mu \,\|\, \nu^{S_1,\ldots,S_k}\right)\right] \leq \frac{2^{\lceil \log_2 s_{\max}\rceil} - 1}{n}\sum_{i=1}^{n} I\left(X_i; \{X_j\}_{j\neq i}\right) = \frac{2^{\lceil \log_2 s_{\max}\rceil} - 1}{n}(\mathrm{TC} + \mathrm{DTC})$$

We provide a short proof of this result via Theorem 16.

Proof The equality in the theorem follows from the definition of TC and DTC, so we aim to prove the inequality. By Theorem 16 and Lemma 8, we have

$$\begin{aligned}\mathbb{E}_{S_1,\ldots,S_k}[\mathsf{KL}(\mu \,\|\, \nu^{S_1,\ldots,S_k})] &= \sum_{i=1}^{k}\left(\sum_{j=1}^{s_i}(Z_{N_{i-1}+j} - Z_{N_{i-1}+1})\right)\\ &\le \sum_{i=1}^{k}(s_i - 1)(Z_{N_i} - Z_{N_{i-1}})\\ &\le (s_{\max} - 1)Z_n\\ &\le \frac{2^{\lceil \log_2 s_{\max}\rceil} - 1}{n}(\mathrm{TC} + \mathrm{DTC}),\end{aligned}$$

where in the second line we have noted that the summand $Z_{N_{i-1}+j} - Z_{N_{i-1}+1}$ is zero for $j = 1$ and at most $Z_{N_i} - Z_{N_{i-1}}$ otherwise. This completes the proof. ■

Remark 32 *We can restate Theorem 31 in terms of query complexity given a fixed ε. In particular, the number of queries is $k \ge \frac{n}{s_{\max}}$. Thus, we find that given any ε, there is a mask schedule attaining expected sampling error at most ε in $\mathcal{O}\left(\left\lceil \frac{\mathrm{TC}+\mathrm{DTC}}{\varepsilon}\right\rceil\right)$ queries.*

As shown in Li and Cai (2025), this upper bound is optimal in some cases. However, as we remarked in the introduction, it is generally worse than Theorem 9.

F.2. Recovering the bound of Austin (Austin (2020))

In Austin (2020), the author proves that distributions with low TC can be decomposed into a fixed subset of coordinates S and a remaining subset of coordinates $[n]\backslash S$ which have low conditional TC. For completeness, we show the result here.

Lemma 33 (Lemma 8.3 of Austin (2020)) *Let μ be the data distribution. Then there is a subset size $s \le \frac{\mathrm{DTC}}{\delta^2}$ for which in expectation over all $|S| = s$, $S \subseteq [n]$, we have*

$$\mathrm{TC}(X_1, \ldots, X_n \mid X_S) + \mathrm{DTC}(X_1, \ldots, X_n \mid X_S) \le \delta^2(n - |S|),$$

where $\mathrm{TC}(Y \mid X) = \mathbb{E}_{x\sim p(X)}\mathrm{TC}(Y \mid X = x)$ is the conditional TC and similarly for $\mathrm{DTC}(Y \mid X)$.

This lemma yields a natural method of sampling μ: first perfectly sample an arbitrary subset of size s, and then sample the remaining coordinates in one-shot.

Corollary 34 *Suppose $\mathrm{DTC} \le \delta^2 n$. Let μ be the data distribution and ν be the output of the random unmasking algorithm $\mathcal{A}(k, \{s_i\}_1^k)$. Then there exists a schedule $(k, \{s_i\}_1^k)$ for which the query complexity satisfies $k \le \frac{\mathrm{DTC}}{\delta^2} + 1$ and the error is bounded by*

$$\mathsf{KL}(\mu \,\|\, \nu) \le \delta^2(n - k + 1).$$

Note that here $k = s + 1$, since there is the final one-shot step. We provide a short proof of this result via Theorem 16.

Proof Consider the schedule given by $k = \left\lfloor \frac{\mathrm{DTC}}{\delta^2}\right\rfloor + 1$, $s_i = 1$ for $i \le k - 1$, and $s_k = n - k + 1$. By Theorem 16, we have that

$$\mathsf{KL}(\mu \,\|\, \nu) = (n - k + 1)(Z_n - Z_k)$$

$$
\begin{aligned}
&\leq \frac{n-k+1}{k}\sum_{j=1}^{k}(Z_n - Z_j) \\
&\leq \frac{n-k+1}{k}\mathrm{DTC} \\
&\leq \delta^2(n-k+1),
\end{aligned}
$$

where we have used Lemma 8 repeatedly. ■

Note that the bound in Corollary 34 is not particularly strong; in particular, the sampling procedure is essentially two-step and does not provide significant flexibility to the choice of mask schedule. This can be improved by replacing the one-shot sample of S_k with an ℓ-step, constant mask size sampler. Under this regime, combining the above result and Theorem 31 below and then optimizing over k and ℓ, we can recover Theorem 10 given in the introduction. We provide a detailed proof below.

Proof [Proof of Theorem 10] Consider the schedule $s_i = 1$ for $i \leq k-1$ and $s_i = \lfloor \frac{n-k+1}{\ell} \rfloor$ for $k \leq i \leq k+\ell+1$.[9] Applying Theorem 31 to the conditional distribution $X_1, \ldots, X_n \mid X_S$ and then Lemma 33, we find that the total error is bounded by

$$
\mathbb{E}\left[\mathsf{KL}(\mu \,\|\, \nu)\right] \leq \frac{1}{\ell}(\mathrm{TC}(X_1, \ldots, X_n \mid X_S) + \mathrm{DTC}(X_1, \ldots, X_n \mid X_S)) \leq \frac{\delta^2 n}{\ell}.
$$

The total query complexity is $k+\ell$. Take $\ell = \left\lceil \frac{\delta^2 n}{\varepsilon} \right\rceil$ and $k \leq \frac{\mathrm{DTC}}{\delta^2}$, observe that we can then set $\delta^2 = \sqrt{\frac{\mathrm{DTC}\cdot\varepsilon}{n}}$. We finally find that this schedule yields error $\mathbb{E}\left[\mathsf{KL}(\mu \,\|\, \nu)\right] \leq \varepsilon$ and query complexity $k+\ell = \mathcal{O}\left(\sqrt{\frac{\mathrm{DTC}\cdot n}{\varepsilon}}\right)$, as desired. This completes the proof. ■

Appendix G. Decoupling estimation error and sampling error

This appendix follows the work of Li and Cai (2025), and is written in order that the present work be self-contained. For simplicity, let $\mu(x)$ denote the PDF of μ. Let $T_i = \cup_{j<i} S_j$ for any subsets $S = \bigsqcup_{i=1}^{k} S_i$. We consider learning the estimate $\widehat{\mathsf{CO}}$ which minimizes the following error:

$$
\mathrm{error}(\mu, \widehat{\mathsf{CO}}) = \mathbb{E}_{S_1,\ldots,S_k; i\in[k]}\left[\frac{n}{|S_i|}\sum_{j\in S_i}\log\frac{\mu(X_j \mid X_{T_i} = x_{T_i})}{\widehat{\mathsf{CO}}(X_j \mid X_{T_i} = x_{T_i})}\right]
$$

where $\widehat{\mathsf{CO}}(X_j \mid X_{T_i} = x_{T_i})$ denotes the conditional marginal of X_j outputted by the learned oracle, and the expectation is over *all* schedules $S = \bigsqcup_{i=1}^{k} S_i$, and i is drawn from the distribution $p(i) = \frac{|S_i|}{n}$.

As a brief remark, note that since $\widehat{\mathsf{CO}}$ only appears in the denominator of the logarithmic term, we can estimate the learning error

$$
\text{learning error} = -\mathbb{E}_{S_1,\ldots,S_k; i\in[k]}\left[\frac{n}{|S_i|}\sum_{j\in S_i}\log\widehat{\mathsf{CO}}(X_j \mid X_{T_i} = x_{T_i})\right]
$$

via training samples; moreover, by positivity of KL, the optimum is precisely $\widehat{\mathsf{CO}} = \mathsf{CO}$, hence with sufficiently many samples, we can expect to learn a good estimate $\widehat{\mathsf{CO}}$.

9. This is approximate, as we do not necessarily have $\sum s_i = n$. Formally, if $\sum_i s_i > n$, omit as many terms s_i from the end as necessary and cap the final value of s_i

Now, the following error-decoupling result justifies our assumption that the conditional oracle $\widehat{\mathsf{CO}}$ is perfect.

Lemma 35 (Li and Cai (2025)) *Let μ be the data distribution and $(S_1, \ldots, S_k)$ be an unmasking schedule. Moreover, let $\widehat{\mathsf{CO}}$ be a learned conditional marginal oracle, which estimates CO. Let $\nu^{S_1,\ldots,S_k}$ be the distribution sampled by $\mathcal{A}_{\text{fixed}}(k, \{S_1\}_1^k)$ using CO, and $\hat{\nu}^{S_1,\ldots,S_k}$ is the distribution sampled by $\mathcal{A}_{\text{fixed}}(k, \{S_1\}_1^k)$ using $\widehat{\mathsf{CO}}$. Then*

$$\mathsf{KL}\big(\mu \,\big\|\, \nu^{S_1,\ldots,S_k})\big) = \mathsf{KL}\big(\mu \,\big\|\, \hat{\nu}^{S_1,\ldots,S_k}\big) + \text{error}(\mu, \widehat{\mathsf{CO}})$$

Proof Observe that

$$\begin{aligned}
\mathsf{KL}\big(\mu \,\big\|\, \nu^{S_1,\ldots,S_k}\big) - \mathsf{KL}\big(\mu \,\big\|\, \hat{\nu}^{S_1,\ldots,S_k}\big) &= \int_{\Sigma^n} \mu(x) \log\left(\frac{\nu^{S_1,\ldots,S_k}(x)}{\hat{\nu}^{S_1,\ldots,S_k}(x)}\right) \mathrm{d}x \\
&= \sum_{i=1}^{k} \int_{\Sigma^n} \mu(x) \log\left(\frac{\nu^{S_1,\ldots,S_k}(x_{S_i} \mid X_{T_i} = x_{T_i})}{\hat{\nu}^{S_1,\ldots,S_k}(x_{S_i} \mid X_{T_i} = x_{T_i})}\right) \mathrm{d}x \\
&= \sum_{i=1}^{k} \sum_{j \in S_i} \log \frac{\mu(X_j \mid X_{T_i} = x_{T_i})}{\widehat{\mathsf{CO}}(X_j \mid X_{T_i} = x_{T_i})} \\
&= \mathbb{E}_{S_1,\ldots,S_k; i \in [k]}\left[\frac{n}{|S_i|} \sum_{j \in S_i} \log \frac{\mu(X_j \mid X_{T_i} = x_{T_i})}{\widehat{\mathsf{CO}}(X_j \mid X_{T_i} = x_{T_i})}\right] \\
&= \text{error}(\mu, \widehat{\mathsf{CO}}),
\end{aligned}$$

where the factor of $\frac{n}{|S_i|}$ is due to the distribution of i in the expectation formula for $\text{error}(\mu, \widehat{\mathsf{CO}})$. ■