UNIFORM DISCRETE DIFFUSION MODELS ARE MINIMAX OPTIMAL FOR ESTIMATING DISTRIBUTIONS WITH SMALL EFFECTIVE SUPPORT SIZE

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete diffusion models have emerged as a practically successful framework for generative modeling on discrete product spaces, yet their statistical generalization properties remain poorly understood. Discrete real-world data such as text or biological sequences often concentrate on a small fraction of the astronomically large ambient space because of semantic or physical constraints, but existing bounds fail to capture this distributional structure and instead scale with the size of the ambient space, giving rise to almost vacuous error bounds. We address this gap for uniform discrete diffusion, one of the two dominant discrete diffusion paradigms alongside masking diffusion, by deriving statistical guarantees governed by the *effective support size*, a sample-size-dependent measure of distributional complexity. Given n independent and identically distributed (i.i.d.) samples from an unknown data distribution P_0 on $[K]^d$, we show that, with appropriate choices of network size and hyperparameters, the expected total variation (TV) loss scales as $\mathcal{O}\left(\sqrt{\mathfrak{s}_n(P_0)/n}\right)$, while the expected Kullback–Leibler (KL) divergence is bounded by $\mathcal{O}\left(\frac{1}{n}\mathfrak{s}_n(P_0)\log(eK^d/\mathfrak{s}_n(P_0))\log n\right)$. Furthermore, we show that the TV rate is minimax optimal and that the KL rate is minimax optimal up to a factor of $\log n$. Together, these upper and lower bounds show that uniform discrete diffusion successfully avoids the curse of dimensionality for distributions with small effective support size: the TV error rate depends on the ambient state-space size only through $\mathfrak{s}_n(P_0)$, while the corresponding KL rate incurs only an additional logarithmic dependence on the ambient state-space size.

1 INTRODUCTION

Diffusion models have become a central paradigm for generative modeling (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021). Extensions to discrete state spaces have carried this paradigm to categorical domains such as text, graphs, molecules, and combinatorial optimization (He et al., 2023; Hoogeboom et al., 2021; Vignac et al., 2023; Sun & Yang, 2023). Austin et al. (2021) introduced structured categorical corruption kernels in discrete time, and Campbell et al. (2022) cast both the forward and reverse processes as continuous-time Markov chains (CTMCs). Score-based approaches that estimate probability ratios through score entropy have since achieved strong empirical performance in language modeling (Lou et al., 2024). Two forward processes dominate: *uniform diffusion*, in which each coordinate evolves toward uniformity, and *masking diffusion*, in which each coordinate is absorbed into a special mask state. They induce distinct reverse-time dynamics and hence different statistical and algorithmic properties (Dmitriev et al., 2026; Zhang et al., 2026). In this work, we focus on coordinatewise uniform diffusion on the product space $[K]^d$.

The empirical success of discrete diffusion has spurred substantial theoretical work. Most of this work concerns the convergence of reverse-time samplers (Campbell et al., 2022; Chen & Ying, 2025; Ren et al., 2025; Zhang et al., 2025; Pham et al., 2025; Liang et al., 2025; Dmitriev et al., 2026), particularly fixed-grid τ-leaping and its variants. For uniform discrete diffusion, representative results (Liang et al., 2025; Dmitriev et al., 2026) take the form

$$D_{\mathrm{KL}}(P_\delta \| Q_\delta) \lesssim \varepsilon_{\mathrm{score}} + e^{-T} d \log K + \mathcal{R}_{\mathrm{disc}}.$$

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

Here, P_δ is the forward marginal at the early-stopping time δ, and Q_δ is the output distribution of the sampler initialized from the uniform distribution at time T. The first term $\varepsilon_{\text{score}}$ bounds the interval-weighted score-entropy error on the sampling grid, and the remaining terms account for initialization and time discretization. These guarantees take an $\varepsilon_{\text{score}}$-accurate score estimate as given and do not establish that such an estimate can be learned. It therefore remains unclear how they translate into finite-sample bounds on the distribution-estimation risk when the score is learned from n training samples.

Without exploiting structure in the data distribution, finite-sample guarantees face a fundamental obstacle: the size of the ambient space. For arbitrary distributions on $[K]^d$, the minimax expected risk is of order $\Theta\left(\min\{1, \sqrt{K^d/n}\}\right)$ under TV loss (Han et al., 2015; Kamath et al., 2015) and $\Theta\left(\log(1 + K^d/n)\right)$ under KL loss (Ye et al., 2025; Mourtada, 2026). Consistent estimation over the unrestricted class therefore requires $K^d/n \to 0$, a condition far from realistic. The Score Entropy Discrete Diffusion (SEDD) language experiments (Lou et al., 2024), for instance, use a vocabulary size $K = 50{,}257$ and sequence length $d = 1024$. The resulting ambient space of size $50{,}257^{1024}$ is astronomically larger than any feasible training set.

The finite-sample statistical theory of discrete diffusion that has recently begun to emerge does not escape this barrier (Wakasugi & Suzuki, 2025; Cho & Wu, 2026; Srikanth et al., 2026; Zhang et al., 2026). The KL bounds of Wakasugi & Suzuki (2025) and Cho & Wu (2026) are of order $\widetilde{\mathcal{O}}(K^d/n)$, and the TV bounds of Zhang et al. (2026) grow exponentially in d, so all become vacuous when $n \ll K^d$. Refinements that avoid this dependence either rely on assumptions on P_0 that are unverifiable in practice, such as fast spectral decay (Wakasugi & Suzuki, 2025), or depend on quantities that can themselves be as large as K^d. For example, the bound of Srikanth et al. (2026) scales polynomially with $\max_{d_{\text{Ham}}(\boldsymbol{x},\boldsymbol{y})=1} P_0(\boldsymbol{y})/P_0(\boldsymbol{x})$, which is enormous when a single-token substitution sharply reduces sequence probability.

Such pessimistic guarantees are at odds with the structure of real data. Text and biological sequences concentrate on a vanishing fraction of $[K]^d$ because of semantic or physical constraints, so the ambient cardinality greatly overstates their statistical complexity. A natural first refinement replaces K^d by the support size $|\operatorname{supp}(P_0)|$. However, the support can itself be comparable to K^d even when most of the mass lies on far fewer states. Moreover, at sample size n, states with probability well below $1/n$ are unlikely to be observed and should not each count as a full unit of complexity.

We therefore measure complexity by the *effective support size* $\mathfrak{s}_n(P_0) := \sum_{\boldsymbol{x}\in[K]^d} \min\{nP_0(\boldsymbol{x}), 1\}$. Each state with probability at least $1/n$ counts as one effective state, while rarer states contribute according to their expected counts in n samples. Up to universal constants, $\mathfrak{s}_n(P_0)$ is the expected number of distinct states in a sample of size n (Lemma A.2). It satisfies $\mathfrak{s}_n(P_0) \le \min\{n, |\operatorname{supp}(P_0)|\}$ and can be far smaller than both. Under Zipfian decay $p_{(j)} \asymp j^{-\alpha}$, for instance, $\mathfrak{s}_n(P_0) \asymp \min\{|\operatorname{supp}(P_0)|, n^{1/\alpha}\}$ (Section 2.1). The notion of effective support size has been used in the classical literature on occupancy counts, unseen species, and missing mass (Good, 1953; Karlin, 1967; Ben-Hamou et al., 2017). Related occupancy-based notions have also yielded distribution-dependent guarantees for discrete distribution estimation (Falahatgar et al., 2017; Mourtada, 2026).

This leads to the central question of the paper:

> *Can discrete diffusion models attain distribution-estimation rates that adapt to $\mathfrak{s}_n(P_0)$ and remain informative even when $n \ll K^d$?*

Summary of contributions. We answer this question affirmatively for uniform discrete diffusion.

- **Adaptive guarantees for uniform discrete diffusion.** We establish end-to-end finite-sample guarantees for uniform discrete diffusion with a learned score. They account for score estimation with ReLU networks as well as initialization and time-discretization errors, and they apply to any fixed-grid sampler satisfying a standard score-entropy guarantee, including τ-leaping. The expected TV loss is $\mathcal{O}\big(\sqrt{\mathfrak{s}_n(P_0)/n}\big)$, with number of sampling steps and network depth, width, and sparsity polynomial in K, d, and $n/\mathfrak{s}_n(P_0)$. With early stopping at $\delta = 1/n$, the expected KL divergence is $\mathcal{O}\big(\frac{1}{n}\mathfrak{s}_n(P_0)\log(eK^d/\mathfrak{s}_n(P_0))\log n\big)$. Both bounds depend on K^d at most logarithmically and remain informative when $n \ll K^d$. The KL guarantee carries particular

weight: the empirical distribution has infinite KL risk whenever some states go unobserved, so attaining it requires generalizing beyond the training samples.

- **Matching minimax lower bounds.** For $2 \le s_n \le \min\{n, K^d\}$, let $\mathcal{P}_{s_n} := \{P_0 \in \Delta_{K^d} : \mathfrak{s}_n(P_0) \le s_n\}$. We establish the non-asymptotic minimax rates

$$\inf_{\widetilde{P}} \sup_{P_0 \in \mathcal{P}_{s_n}} \mathbb{E}_{P_0} d_{\mathrm{TV}}(P_0, \widetilde{P}) \asymp \sqrt{\frac{s_n}{n}}, \qquad \inf_{\widetilde{P}} \sup_{P_0 \in \mathcal{P}_{s_n}} \mathbb{E}_{P_0} D_{\mathrm{KL}}(P_0 \| \widetilde{P}) \asymp \frac{s_n \log(eK^d/s_n)}{n}.$$

Hence, uniform discrete diffusion is minimax optimal under TV loss and optimal up to a $\log n$ factor under KL loss. When $s_n = \min\{n, K^d\}$, $\mathcal{P}_{s_n} = \Delta_{K^d}$ holds, and our rates reduce to the classical minimax rates for arbitrary distributions on K^d states. Thus, our result extends the classical unrestricted theory, while yielding potentially much sharper rates for distributions with small effective support size. Allowing K and d to vary with n, these rates also characterize when consistent estimation is possible. Under TV loss, this holds if and only if $s_n/n \to 0$, and under KL loss, if and only if $(s_n/n)\log(eK^d/s_n) \to 0$. The gap between the two conditions yields three regimes (Figure 1): consistent estimation is impossible, possible under TV loss only, or possible under both losses.

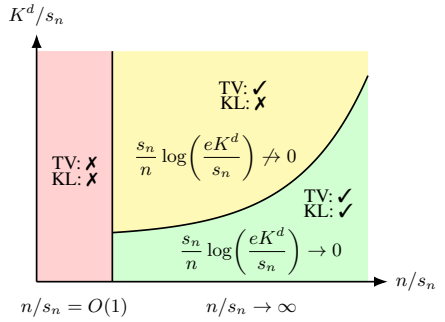


Figure 1: Asymptotic statistical regimes for estimating discrete distributions on $[K]^d$ with n i.i.d. observations.

2 Background and Problem Setting

Notations. We use plain uppercase letters such as P and Q for probability distributions and bold uppercase letters such as $\boldsymbol{P}$ and $\boldsymbol{Q}$ for transition and generator matrices. For a positive integer K, write $[K] := \{1, \dots, K\}$, and let Δ_{K^d} denote the probability simplex on $[K]^d$. For a probability distribution P on $[K]^d$, let $\mathrm{supp}(P) := \{\boldsymbol{x} : P(\boldsymbol{x}) > 0\}$ denote its support, and write $\mathbb{E}_P$ for expectation under P. For probability distributions P and Q on $[K]^d$, the Total Variation (TV) distance and Kullback-Leibler divergence are defined as, $d_{\mathrm{TV}}(P, Q) := \frac{1}{2}\sum_{\boldsymbol{x}\in[K]^d} |P(\boldsymbol{x}) - Q(\boldsymbol{x})|$, $D_{\mathrm{KL}}(P\|Q) := \sum_{\boldsymbol{x}\in[K]^d} P(\boldsymbol{x}) \log\left(P(\boldsymbol{x})/Q(\boldsymbol{x})\right)$, where, we use the conventions $0\log(0/q) := 0$ and $p\log(p/0) := \infty$ for $p > 0$. For nonnegative quantities u_n and v_n, which are functions of n, we write, $u \lesssim v$ if $u_n \le C v_n$ for a universal constant $C > 0$, independent of n. We write $u_n \asymp v_n$ if both $u_n \lesssim v_n$ and $v_n \lesssim u_n$. For $\boldsymbol{x} = (x^1, \dots, x^d), \boldsymbol{y} = (y^1, \dots, y^d) \in [K]^d$, the Hamming distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as, $d_{\mathrm{Ham}}(\boldsymbol{x}, \boldsymbol{y}) := \sum_{i=1}^d \mathbf{1}\{x^i \ne y^i\}$, where $\mathbf{1}(\cdot)$ denotes the indicator function. For $u, v > 0$, define the Bregman divergence generated by $f(u) := u\log u$ as $D_{\mathrm{Br}}(u, v) := f(u) - f(v) - f'(v)(u - v) = u\log(u/v) - u + v$. For $a \in [K]$, let $\boldsymbol{e}_a \in \{0, 1\}^K$ be the a-th standard basis vector, and define the coordinatewise one-hot encoding $\boldsymbol{e}(\boldsymbol{x}) := (\boldsymbol{e}_{x^1}^\top, \dots, \boldsymbol{e}_{x^d}^\top)^\top \in \{0, 1\}^{dK}$. $\delta_{\boldsymbol{x}}$ denotes the Dirac distribution at $\boldsymbol{x}$. For $\ell \le r$, define $\mathrm{clip}_{[\ell,r]}(u) := \min\{\max\{u, \ell\}, r\}$. All logarithms are with respect to (w.r.t.) the natural base.

2.1 Statistical Setup and Effective Support Size

We consider distribution estimation on the product space $[K]^d$ for $K, d \ge 2$, where K can be interpreted as the vocabulary size and d is the sequence length. Let $P_0 \in \Delta_{K^d}$ be an unknown data distribution, and suppose that we observe training n i.i.d. samples $X_0^{(1)}, \dots, X_0^{(n)} \overset{\text{i.i.d.}}{\sim} P_0$ for $n \ge 2$. A distribution estimator $\widetilde{P} \in \Delta_{K^d}$ is any (possibly randomized) data-dependent estimator of P_0. We study its expected total variation and Kullback–Leibler risks, $\mathbb{E}_{P_0} d_{\mathrm{TV}}(P_0, \widetilde{P})$ and $\mathbb{E}_{P_0} D_{\mathrm{KL}}(P_0\|\widetilde{P})$.

Effective support size. As a measure of the complexity of P_0 relative to a sample of size n, we use its *effective support size* (Mourtada, 2026; Falahatgar et al., 2017),

$$\mathfrak{s}_n(P_0) := \sum_{\boldsymbol{x}\in[K]^d} \min\{nP_0(\boldsymbol{x}), 1\} = \left|\left\{\boldsymbol{x} \in [K]^d : P_0(\boldsymbol{x}) \ge 1/n\right\}\right| + n \sum_{\boldsymbol{x} \,:\, P_0(\boldsymbol{x}) < 1/n} P_0(\boldsymbol{x}),$$

where states above the sampling resolution $1/n$ each contribute a unit, while states below it contribute only their rescaled aggregate mass. Thus $\mathfrak{s}_n(P_0)$ is nondecreasing in n, satisfies $\mathfrak{s}_n(P_0) \le \min\{n, |\operatorname{supp}(P_0)|\}$, and recovers the exact support size as $n \to \infty$, i.e., $\lim_{n\to\infty} \mathfrak{s}_n(P_0) = \mathfrak{s}_\infty(P_0) := \left|\left\{\boldsymbol{x} \in [K]^d : P_0(\boldsymbol{x}) > 0\right\}\right|$. This refines the crude sparsity measure $|\operatorname{supp}(P_0)|$ into one calibrated to the sample size, and is sensitive not just to $|\operatorname{supp}(P_0)|$ but to the decay of the point masses. Writing $p_{(1)} \ge \cdots \ge p_{(|\operatorname{supp}(P_0)|)} > 0$ for the sorted nonzero masses of P_0, Lemma A.1 gives, for $n \ge 2$,

$$\mathfrak{s}_n(P_0) \asymp \begin{cases} \min\{|\operatorname{supp}(P_0)|, n\}, & \text{if } p_{(j)} \asymp |\operatorname{supp}(P_0)|^{-1}, \\ \min\{|\operatorname{supp}(P_0)|, n^{1/\alpha}\}, & \text{if } p_{(j)} \asymp j^{-\alpha}, \\ \min\{|\operatorname{supp}(P_0)|, \log n\}, & \text{if } p_{(j)} \asymp e^{-cj}, \end{cases} \tag{1}$$

uniformly over $1 \le j \le |\operatorname{supp}(P_0)|$, with $\alpha > 1$, $c > 0$ fixed and implicit constants independent of n and $|\operatorname{supp}(P_0)|$. These regimes span the full range of behavior: linear growth under near-uniform mass, sublinear polynomial growth $n^{1/\alpha}$ under Zipfian decay, and merely logarithmic growth under geometric decay. It is this decay-dependent behavior, rather than the ambient size K^d or the raw support size $|\operatorname{supp}(P_0)|$, that governs the finite-sample rates established below.

For $1 \le s_n \le \min\{n, K^d\}$, we define the effective support class $\mathcal{P}_{s_n} = \mathcal{P}_{s_n}(K^d, n) := \{P_0 \in \Delta_{K^d} : \mathfrak{s}_n(P_0) \le s_n\}$. The corresponding minimax risks in the TV and KL are

$$\inf_{\widetilde{P}} \sup_{P_0 \in \mathcal{P}_{s_n}} \mathbb{E}_{P_0} d_{\mathrm{TV}}(P_0, \widetilde{P}) \qquad \text{and} \qquad \inf_{\widetilde{P}} \sup_{P_0 \in \mathcal{P}_{s_n}} \mathbb{E}_{P_0} D_{\mathrm{KL}}(P_0 \| \widetilde{P}).$$

The goal of this paper is to study the behavior of the above minimax risks and to understand under what model and hyperparameter choices discrete diffusion models can achieve this optimal rate.

2.2 Uniform Discrete Diffusion

The uniform discrete diffusion consists of a forward noising process, its score-based time reversal denoising, a fixed-grid score estimator, and a reverse sampler.

Forward process. The forward process is a continuous-time Markov chain (CTMC) on $[K]^d$ with independently evolving coordinates. The time-homogeneous generator of each coordinate is $\boldsymbol{Q}^{\mathrm{coord}}(a, b) := 1/K - \mathbf{1}\{a = b\}$ for every $a, b \in [K]$, which means that each coordinate jumps to another state with rate $1/K$. Following Campbell et al. (2022), the induced generator $\boldsymbol{Q}$ on $[K]^d$ is

$$\boldsymbol{Q}(\boldsymbol{x}, \boldsymbol{y}) = K^{-1}\mathbf{1}\left\{d_{\mathrm{Ham}}(\boldsymbol{x}, \boldsymbol{y}) = 1\right\} - dK^{-1}(K-1)\mathbf{1}\left\{\boldsymbol{x} = \boldsymbol{y}\right\}.$$

The one-coordinate transition kernel is given by $\boldsymbol{P}_t^{\mathrm{coord}}(a, b) = e^{-t}\mathbf{1}\{a = b\} + (1 - e^{-t})/K$, and coordinatewise independence gives $\boldsymbol{P}_t(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^d \boldsymbol{P}_t^{\mathrm{coord}}(x^i, y^i)$ and $\boldsymbol{P}_t = e^{t\boldsymbol{Q}}$. Writing $P_t := P_0\boldsymbol{P}_t$, symmetry of $\boldsymbol{Q}$ implies that the uniform distribution $\pi(\boldsymbol{x}) := K^{-d}$ is stationary. Since the chain is irreducible on the finite state space $[K]^d$, π is the unique stationary distribution and $P_t \to \pi$ as $t \to \infty$ (Norris, 1997, Theorem 3.6.2).

Reverse process and discrete score. Fix a terminal time $T > 0$, and let $u \in [0, T]$ denote reverse time. Initialized from P_T, the time reversal of the forward process is a time-inhomogeneous CTMC whose off-diagonal rates, for $u \in [0, T)$, are $\overleftarrow{\boldsymbol{Q}}_u^P(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{Q}(\boldsymbol{y}, \boldsymbol{x})\frac{P_{T-u}(\boldsymbol{y})}{P_{T-u}(\boldsymbol{x})}$, $\boldsymbol{x} \ne \boldsymbol{y}$. The diagonal entries are chosen so that each row sums to zero. With the forward generator fixed, the probability ratio $P_{T-u}(\boldsymbol{y})/P_{T-u}(\boldsymbol{x})$ determines the reverse generator. This motivates defining the true discrete score, for $t > 0$, as $\sigma_t^\star(\boldsymbol{y}, \boldsymbol{x}) := \frac{P_t(\boldsymbol{y})}{P_t(\boldsymbol{x})}$. This ratio plays a role analogous to the continuous score $\nabla \log p_t$, which governs the reverse drift in continuous diffusion models. For our coordinatewise uniform forward process, the off-diagonal reverse rates become

$$\overleftarrow{\boldsymbol{Q}}_u^P(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} K^{-1}\sigma_{T-u}^\star(\boldsymbol{y}, \boldsymbol{x}), & d_{\mathrm{Ham}}(\boldsymbol{x}, \boldsymbol{y}) = 1, \\ 0, & d_{\mathrm{Ham}}(\boldsymbol{x}, \boldsymbol{y}) \ge 2, \end{cases} \qquad \boldsymbol{x} \ne \boldsymbol{y}.$$

For sufficiently large T, generation starts from the tractable stationary distribution π in place of P_T and simulates the reverse process using a learned approximation of $\sigma_t^\star$.

Fixed-grid score estimation. Although the reverse process evolves continuously in time, a practical sampler evaluates the score at only finitely many grid-points. In this paper, we choose a deterministic forward-time grid $\mathcal{T} := \{\delta = t_0 < t_1 < \cdots < t_N = T\}$, where $\delta \geq 0$ is the early-stopping time and scores are estimated only at $t_1, \ldots, t_N$. For a positive candidate score σ_t, the score-entropy loss at time $t > 0$ is given by,

$$\mathcal{L}_t^{\mathrm{SE}}(\sigma; P_0) := \mathbb{E}_{X_t \sim P_t}\left[\sum_{\boldsymbol{y}:\boldsymbol{Q}(\boldsymbol{y},X_t)>0} \boldsymbol{Q}(\boldsymbol{y}, X_t) D_{\mathrm{Br}}\left(\sigma_t^\star(\boldsymbol{y}, X_t), \sigma_t(\boldsymbol{y}, X_t)\right)\right].$$

This loss is nonnegative and minimized by the true score $\sigma_t^\star$ on all pairs $(\boldsymbol{y}, \boldsymbol{x})$ satisfying $d_{\mathrm{Ham}}(\boldsymbol{y}, \boldsymbol{x}) = 1$ (Lou et al., 2024). We aggregate the score error over the grid using the interval lengths as weights: $\mathcal{L}_{\mathcal{T}}^{\mathrm{SE}}(\sigma; P_0) := \sum_{k=1}^{N}(t_k - t_{k-1})\mathcal{L}_{t_k}^{\mathrm{SE}}(\sigma; P_0)$. Since $\mathcal{L}_t^{\mathrm{SE}}$ depends on the intractable marginal P_t, we instead use its denoising form. Denoting the true conditional score by $\sigma_{t|0}^\star(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{x}_0) := \boldsymbol{P}_t(\boldsymbol{x}_0, \boldsymbol{y})/\boldsymbol{P}_t(\boldsymbol{x}_0, \boldsymbol{x})$, the denoising score-entropy loss is constructed as:

$$\mathcal{L}_t^{\mathrm{DSE}}(\sigma; P_0) := \quad \mathbb{E}\left[\sum_{\boldsymbol{y}:\boldsymbol{Q}(\boldsymbol{y},X_t)>0} \boldsymbol{Q}(\boldsymbol{y}, X_t)\left(\sigma_t(\boldsymbol{y}, X_t) - \sigma_{t|0}^\star(\boldsymbol{y}, X_t \mid X_0)\log \sigma_t(\boldsymbol{y}, X_t)\right)\right],$$

where in the above expectation, $X_0 \sim P_0$, and $X_t | X_0 \sim \boldsymbol{P}_t(X_0, \cdot)$. As shown in Lemma A.14, the score-entropy and denoising score-entropy losses differ only by a term independent of σ and therefore have the same minimizer. Accordingly, we define the denoising score-matching loss as $\mathcal{L}_{\mathcal{T}}^{\mathrm{DSE}}(\sigma; P_0) := \sum_{k=1}^{N}(t_k - t_{k-1})\mathcal{L}_{t_k}^{\mathrm{DSE}}(\sigma; P_0)$. During training, since P_0 is unknown, it is substituted by the empirical distribution $\widehat{P}_0 := n^{-1}\sum_{i=1}^{n}\delta_{X_0^{(i)}}$, yielding the empirical training objective

$$\widehat{\mathcal{L}}_{\mathcal{T}}^{\mathrm{DSE}}(\sigma) := \mathcal{L}_{\mathcal{T}}^{\mathrm{DSE}}(\sigma; \widehat{P}_0). \tag{2}$$

Neural-network score class. For positive integers L, W, S and $B > 0$, let $\mathcal{G}(L, W, S, B)$ denote the class of scalar-output ReLU networks on $\mathbb{R}^{1+2dK}$ with depth at most L, width at most W, at most S nonzero parameters, and all weights and biases bounded in absolute value by B. For $G \in \mathcal{G}(L, W, S, B)$ and $M \geq 1$, define

$$\sigma_{t_k}^G(\boldsymbol{y}, \boldsymbol{x}) := \mathrm{clip}_{[M^{-1}, M]}\left(G\left(k/N, \boldsymbol{e}(\boldsymbol{y}), \boldsymbol{e}(\boldsymbol{x})\right)\right), \qquad k \in [N], \quad \boldsymbol{x}, \boldsymbol{y} \in [K]^d.$$

Here, k/N encodes only the grid index and does not represent the physical diffusion time t_k. Although $\sigma_{t_k}^G(\boldsymbol{y}, \boldsymbol{x})$ is defined for all $\boldsymbol{x}, \boldsymbol{y} \in [K]^d$, only values with $d_{\mathrm{Ham}}(\boldsymbol{x}, \boldsymbol{y}) = 1$ enter the score-estimation objective and reverse dynamics, since $\boldsymbol{Q}(\boldsymbol{y}, \boldsymbol{x}) = 0$ otherwise. Thus, the network needs only represent the score at the prescribed grid times and on pairs of states at Hamming distance one. Let $\mathcal{F}(L, W, S, B; M) := \{\sigma^G : G \in \mathcal{G}(L, W, S, B)\}$ denote the resulting clipped score class. We assume that the learned score function $\hat{\sigma}$ is an $\varepsilon_{\mathrm{opt}}$-approximate minimizer of the empirical objective (2), i.e.,

$$\widehat{\mathcal{L}}_{\mathcal{T}}^{\mathrm{DSE}}(\widehat{\sigma}) \leq \inf_{\sigma \in \mathcal{F}(L,W,S,B;M)} \widehat{\mathcal{L}}_{\mathcal{T}}^{\mathrm{DSE}}(\sigma) + \varepsilon_{\mathrm{opt}}, \tag{3}$$

where $\varepsilon_{\mathrm{opt}} \geq 0$ is the optimization error.

Reverse sampler. At forward time t, the learned score $\widehat{\sigma}$ provides the estimated reverse rates $\boldsymbol{Q}(\boldsymbol{y}, \boldsymbol{x})\widehat{\sigma}_t(\boldsymbol{y}, \boldsymbol{x})$ for $\boldsymbol{x} \neq \boldsymbol{y}$. Starting from the stationary distribution π at time T, a reverse sampler uses these rate estimates to approximate the reverse-time evolution down to the early-stopping time δ. Examples of fixed-grid samplers include τ-leaping (Gillespie, 2001; Campbell et al., 2022), as well as the Euler method and the Tweedie τ-leaping sampler (Lou et al., 2024). Non-fixed-grid methods include Gillespie-based samplers (Gillespie, 1976; Liu et al., 2025) and uniformization (Jensen, 1953; Chen & Ying, 2025), with score evaluations at random or adaptively chosen times. We focus on fixed-grid samplers, whose score evaluations are restricted to a prescribed deterministic time grid $\mathcal{T}$.

Definition 1 (Fixed-grid score-based sampler)**.** A fixed-grid score-based sampler $\mathcal{A}$ takes as input an initial draw $X_T \sim \pi$, a grid $\mathcal{T}$, and the grid score functions $\{\sigma_{t_k}\}_{k=1}^N$, and returns a random state $X_\delta^{\mathcal{A},\mathcal{T},\sigma} \in [K]^d$. We denote its output distribution by $Q_\delta^{\mathcal{A},\mathcal{T},\sigma}$.

For the fixed grid $\mathcal{T}$ and the learned score $\widehat{\sigma}$ in (3), we abbreviate the output state and distribution as $X_\delta^{\mathcal{A}}$ and $Q_\delta^{\mathcal{A}}$, respectively. For $M \geq 1$, define the class of M-bounded grid scores by $\Sigma_M(\mathcal{T}) := \left\{\sigma : M^{-1} \leq \sigma_{t_k}(\boldsymbol{y}, \boldsymbol{x}) \leq M \text{ for every } k \in [N],\ \boldsymbol{x}, \boldsymbol{y} \in [K]^d\right\}$. We formulate our sampler-agnostic analysis through the following assumption, which separates the score-estimation error from an approximation remainder accounting for initialization and time discretization.

Assumption 2 (Score-based sampler guarantee)**.** For any $P_0 \in \Delta_{K^d}$, grid $\mathcal{T}$, $M \geq 1$, and score $\sigma \in \Sigma_M(\mathcal{T})$, $D_{\mathrm{KL}}\left(P_\delta \| Q_\delta^{\mathcal{A},\mathcal{T},\sigma}\right) \leq C_{\mathcal{A}}\left\{\mathcal{L}_{\mathcal{T}}^{\mathrm{SE}}(\sigma; P_0) + \mathcal{R}_{\mathcal{A}}(\mathcal{T}; M)\right\}$, where $C_{\mathcal{A}} > 0$ depends only on the sampler. The remainder $\mathcal{R}_{\mathcal{A}}(\mathcal{T}; M) \geq 0$ is deterministic and independent of P_0 and of the particular score values beyond the bound M.

Guarantees based on interval-weighted grid score-entropy loss are established for fixed-grid samplers under uniform discrete diffusion in Liang et al. (2025); Dmitriev et al. (2026). In particular, Theorem 1 of Dmitriev et al. (2026), restated in our notation as Theorem A.15, provides a guarantee for τ-leaping under its stated grid conditions, with remainder $\mathcal{R}_\tau(\mathcal{T}; M) = e^{-T} d \log K + \Delta d \log(K/\Delta)$, where Δ denotes the largest grid interval.

3 Statistical Guarantees

3.1 Guarantee in Total Variation

We begin our study of distribution estimation under the TV loss, with proofs appearing in Appendix B. For any fixed-grid sampler satisfying Assumption 2, Theorem 3 shows that uniform discrete diffusion attains the rate $\mathcal{O}(\sqrt{\mathfrak{s}_n(P_0)/n})$ through score estimation and reverse-time sampling. We complement this upper bound with a matching minimax lower bound over $\mathcal{P}_{s_n}$ in Theorem 5. Let $D_n := |\{X_0^{(1)}, \ldots, X_0^{(n)}\}|$ denote the number of distinct observations in the training sample.

Theorem 3 (TV upper bound for uniform discrete diffusion)**.** *Define the natural clipping bound* $M := \frac{1+(K-1)e^{-t_1}}{1-e^{-t_1}}$, *which uniformly bounds the true score between states of Hamming distance one for all* $t \geq t_1$. *Suppose that the fixed-grid sampler* $\mathcal{A}$ *satisfies Assumption 2 and that* $\mathcal{R}_{\mathcal{A}}(\mathcal{T}; M) \lesssim \mathfrak{s}_n(P_0)/n$, $\varepsilon_{\mathrm{opt}} \lesssim \mathfrak{s}_n(P_0)/n$, $d\delta \lesssim \sqrt{\mathfrak{s}_n(P_0)/n}$, *and* $T \geq 1$. *Then there exists a clipped ReLU score class* $\mathcal{F}(L, W, S, B; M)$ *with* $L \lesssim d^2 \log^2\left(\frac{K}{1-e^{-t_1}}\right) + \log^2\left(\frac{nT}{\mathfrak{s}_n(P_0)}\right)$, $W \lesssim dKN + dD_n + d^3 \log^3\left(\frac{K}{1-e^{-t_1}}\right) + \log^3\left(\frac{nT}{\mathfrak{s}_n(P_0)}\right)$, $S \lesssim dK^2N + d^2 D_n \log\left(\frac{K}{1-e^{-t_1}}\right) + dD_n \log\left(\frac{nT}{\mathfrak{s}_n(P_0)}\right) + d^4 \log^4\left(\frac{K}{1-e^{-t_1}}\right) + \log^4\left(\frac{nT}{\mathfrak{s}_n(P_0)}\right)$, $B \lesssim N + \left(\frac{K}{1-e^{-t_1}}\right)^{2d} + \frac{dnMT}{\mathfrak{s}_n(P_0)}$, *such that every* $\varepsilon_{\mathrm{opt}}$*-optimizer* $\widehat{\sigma}$ *of* $\widehat{\mathcal{L}}_{\mathcal{T}}^{\mathrm{DSE}}$ *over this class satisfies*

$$\mathbb{E}_{P_0} d_{\mathrm{TV}}\left(P_0, Q_\delta^{\mathcal{A}}\right) \lesssim \sqrt{\frac{\mathfrak{s}_n(P_0)}{n}}.$$

Theorem 3 is stated for a generic fixed-grid sampler and applies once its remainder term $\mathcal{R}_{\mathcal{A}}$ is controlled at the stated rate. For the τ-leaping sampler, instantiating Theorem 3 with this remainder yields an explicit sampling grid and matching architecture bounds as shown in Corollary 4.

Corollary 4 (TV guarantee for τ-leaping sampler)**.** *For the* τ*-leaping sampler, the grid choice* $\mathcal{T} = \left\{\delta = 0, \Delta, \ldots, (N-1)\Delta, T = \log\left(\frac{edn \log K}{\mathfrak{s}_n(P_0)}\right)\right\}$, *where* $\Delta = \frac{\mathfrak{s}_n(P_0)}{dn \log(Kdn/\mathfrak{s}_n(P_0))}$, *and* $N = \lceil \frac{T}{\Delta} \rceil$ *satisfies the sampler conditions of Theorem 3. The clipped ReLU score class can be chosen with* $M = \frac{1+(K-1)e^{-\Delta}}{1-e^{-\Delta}}$, $L \lesssim d^2 \log^2\left(\frac{Kdn}{\mathfrak{s}_n(P_0)}\right)$, $W \lesssim \frac{d^2 Kn}{\mathfrak{s}_n(P_0)} \log^2\left(\frac{Kdn}{\mathfrak{s}_n(P_0)}\right) + dD_n + d^3 \log^3\left(\frac{Kdn}{\mathfrak{s}_n(P_0)}\right)$, $S \lesssim \frac{d^2K^2n}{\mathfrak{s}_n(P_0)} \log^2\left(\frac{Kdn}{\mathfrak{s}_n(P_0)}\right) + d^2 D_n \log\left(\frac{Kdn}{\mathfrak{s}_n(P_0)}\right) + d^4 \log^4\left(\frac{Kdn}{\mathfrak{s}_n(P_0)}\right)$, *and* $B \lesssim \left[\frac{Kdn}{\mathfrak{s}_n(P_0)} \log\left(\frac{Kdn}{\mathfrak{s}_n(P_0)}\right)\right]^{2d}$.

Interpretation of the sampling and network parameters. The choices of the terminal time T and grid spacing Δ in Corollary 4 calibrate the initialization and discretization errors to the statistical scale $\mathfrak{s}_n(P_0)/n$, while no early stopping is needed under TV loss, so we take $\delta = 0$. The resulting number of score evaluations and clipping level satisfy $N = \widetilde{\mathcal{O}}(dn/\mathfrak{s}_n(P_0))$ and $M = \widetilde{\mathcal{O}}(Kdn/\mathfrak{s}_n(P_0))$, respectively. Together with the displayed bounds on the network depth L, width W, and sparsity S, these quantities depend polynomially on K, d, and $n/\mathfrak{s}_n(P_0)$, rather than on the ambient cardinality K^d. Thus, the TV guarantee avoids polynomial dependence on the ambient state-space size in the number of score evaluations, the clipping level, and the network size. The number of distinct states in the sample, D_n, can be upper bounded by $\mathfrak{s}_n(P_0)$ in expectation and by high probability by Lemma A.2.

Minimax lower bound. As a classical benchmark, Lemma A.6 shows that the empirical distribution $\widehat{P}_0$ also achieves the rate $\sqrt{\mathfrak{s}_n(P_0)/n}$. Theorem 5 shows that this rate is minimax optimal over $\mathcal{P}_{s_n}$.

Theorem 5 (Minimax rate under TV loss). *For every* $2 \le s_n \le \min\{n, K^d\}$,

$$\inf_{\widetilde{P}} \sup_{P_0 \in \mathcal{P}_{s_n}} \mathbb{E}_{P_0} d_{\mathrm{TV}}(P_0, \widetilde{P}) \asymp \sqrt{s_n/n}.$$

Comparison with existing TV bounds. Since the score-entropy loss is naturally aligned with KL divergence, relatively few works derive TV guarantees for uniform discrete diffusion. To our knowledge, only Zhang et al. (2026) obtain a TV bound for uniform discrete diffusion, of order $\widetilde{\mathcal{O}}\Big(\sqrt{dK^{d+1}/n}\Big)$, which becomes vacuous when $n \ll K^d$. In contrast, our rate depends on the effective support size $\mathfrak{s}_n(P_0)$ and, by Theorem 5, attains the minimax-optimal rate over the corresponding effective support classes. In the non-saturated regimes of (1), this yields rates $\mathcal{O}(n^{(\alpha^{-1}-1)/2})$ and $\mathcal{O}(\sqrt{\log n/n})$ for Zipfian and geometric decay, respectively.

3.2 Guarantee in KL Divergence

We now turn to distribution estimation under KL loss, with proofs appearing in Appendix C. Unlike in the TV setting, the empirical distribution $\widehat{P}_0$ generally has infinite expected KL risk. Thus, obtaining a meaningful KL guarantee requires assigning positive probability beyond the observed training samples, making KL loss a stringent measure of distributional generalization. For any fixed-grid sampler satisfying Assumption 2, Theorem 6 shows that uniform discrete diffusion with early-stopping time $\delta = 1/n$ attains the rate $\mathcal{O}\big(\frac{1}{n}\mathfrak{s}_n(P_0)\log(eK^d/\mathfrak{s}_n(P_0))\log n\big)$. We complement this upper bound with a minimax lower bound over $\mathcal{P}_{s_n}$ in Theorem 8, showing that uniform discrete diffusion is minimax optimal up to a factor of $\log n$. Together, these results show that uniform discrete diffusion can generalize well beyond the observed training samples rather than merely memorizing the empirical distribution.

Theorem 6 (KL guarantee for uniform discrete diffusion). *Set early-stopping time* $\delta := 1/n$. *Define the natural clipping bound* $M := \frac{1+(K-1)e^{-t_1}}{1-e^{-t_1}}$, *which uniformly bounds the true score between states of Hamming distance one for all* $t \ge t_1$. *Suppose that the fixed-grid sampler* $\mathcal{A}$ *satisfies Assumption 2 and that* $\mathcal{R}_{\mathcal{A}}(\mathcal{T}; M) \lesssim \left(\frac{1-e^{-1/n}}{K}\right)^d$, $\varepsilon_{\mathrm{opt}} \lesssim \left(\frac{1-e^{-1/n}}{K}\right)^d$, *and* $T \ge 1$. *Then there exists a clipped ReLU score class* $\mathcal{F}(L, W, S, B; M)$ *with* $L \lesssim d^2\log^2(Kn) + \log^2 T$, $W \lesssim dKN + dD_n + d^3\log^3(Kn) + \log^3 T$, $S \lesssim dK^2N + d^2D_n\log(Kn) + dD_n\log T + d^4\log^4(Kn) + \log^4 T$, *and* $B \lesssim N + \left(\frac{K}{1-e^{-t_1}}\right)^{2d} + dMT\left(\frac{K}{1-e^{-1/n}}\right)^d$, *such that every* $\varepsilon_{\mathrm{opt}}$*-optimizer* $\widehat{\sigma}$ *of* $\widehat{\mathcal{L}}^{\mathrm{DSE}}_{\mathcal{T}}$ *over this class satisfies*

$$\mathbb{E}_{P_0} D_{\mathrm{KL}}\left(P_0 \| Q^{\mathcal{A}}_{\delta}\right) \lesssim \frac{1}{n}\mathfrak{s}_n(P_0)\log\big(eK^d/\mathfrak{s}_n(P_0)\big)\log n.$$

Once again, for the τ-leaping sampler, we can obtain specific sampling grid and network architecture choices as highlighted in Corollary 7.

Corollary 7 (KL guarantee for τ-leaping sampler). *For the* τ*-leaping sampler, the grid* $\mathcal{T} = \{\delta = \frac{1}{n}, \frac{1}{n}+\Delta, \ldots, \frac{1}{n}+(N-1)\Delta, T\}$ *with* $T = \log\big[ed(K/(1-e^{-1/n}))^d\log K\big]$, $\Delta = \big(d^2(K/(1-e^{-1/n}))^d\log(Kn)\big)^{-1}$, *and* $N = \lceil (T-1/n)/\Delta \rceil$ *satisfies the sampler conditions of Theorem 6. The clipped ReLU score class can be chosen with* $M = \frac{1+(K-1)e^{-(1/n+\Delta)}}{1-e^{-(1/n+\Delta)}}$, $L \lesssim d^2\log^2(Kn)$, $W \lesssim d^4K(K/(1-e^{-1/n}))^d\log^2(Kn)$, $S \lesssim d^4K^2(K/(1-e^{-1/n}))^d\log^2(Kn)$, *and* $B \lesssim (K/(1-e^{-1/n}))^{2d}$.

Interpretation of the sampling and network parameters. Unlike under TV loss, early stopping is essential under KL loss. We take $\delta = 1/n$ since the early-stopped empirical distribution $\widehat{P}_{1/n}$ already achieves the rate of Theorem 6. The choices of the terminal time T and grid spacing Δ in Corollary 7 calibrate the initialization and discretization errors to the sampler-accuracy scale $\big((1-e^{-1/n})/K\big)^d$, which is much smaller than the final statistical KL rate, unlike in the TV case. This difference arises from the absence of a triangle inequality for KL divergence. Instead, we use Lemma A.13,

which requires the sampling error to be controlled at the uniform lower bound on the probabilities at time $1/n$. Although the clipping level $M = \widetilde{\mathcal{O}}(Kn)$ and network depth L remain mild, the number of sampling steps N, network width W, and sparsity S incur exponential dependence on d. This unfavorable computational dependence is not intrinsic to the statistical KL rate: in Appendix D, we construct an exact one-step empirical reverse sampler that achieves the same statistical rate with only $\mathcal{O}(d(D_n + K))$ operations per generated sample.

Minimax lower bound. The following theorem characterizes the minimax rate over the effective support class $\mathcal{P}_{s_n}$, showing that uniform discrete diffusion achieves the minimax rate up to a factor of $\log n$.

Theorem 8 (Minimax rate under KL loss)**.** *For every* $2 \le s_n \le \min\{n, K^d\}$,

$$\inf_{\widetilde{P}} \sup_{P_0 \in \mathcal{P}_{s_n}} \mathbb{E}_{P_0} D_{\mathrm{KL}}(P_0 \| \widetilde{P}) \asymp \frac{s_n}{n} \log(eK^d/s_n).$$

Comparison with classical distribution estimators. There has been substantial interest in estimating discrete distributions (Braess et al., 2002; Paninski, 2004; Braess & Sauer, 2004; van der Hoeven et al., 2025) and the related missing mass problem (Good, 1953; Orlitsky & Suresh, 2015). Classical asymptotic theory shows that, for a fixed state space of size K^d, the minimax expected KL risk is $(K^d - 1)/(2n) + o(1/n)$ as $n \to \infty$. This fixed state space asymptotic does not describe the large state space regime relevant to discrete diffusion.

Since the empirical distribution $\widehat{P}_0$ generally has infinite KL risk, classical discrete distribution estimation has long relied on smoothing to assign positive probability to unseen states. Among fixed additive-smoothing estimators of the form $(n\widehat{P}_0(x) + \lambda)/(n + \lambda K^d)$, the Laplace estimator with $\lambda = 1$ achieves the minimax-optimal expected risk of order $\Theta\left(\log(1 + K^d/n)\right)$ over the unrestricted class Δ_{K^d} (Paninski, 2004; Mourtada, 2026). However, this rate does not vanish when $n \ll K^d$. The adaptive additive-smoothing estimator with $\lambda = D_n/K^d$ (Mourtada, 2026) and the absolute-discounting estimator (Falahatgar et al., 2017) both achieve the minimax-optimal rate over the effective support class of Theorem 8. Moreover, when the effective support size is maximal, i.e., $s_n = \min\{n, K^d\}$, which corresponds to the unrestricted class, the minimax rates in Theorems 5 and 8 recover the corresponding classical unrestricted minimax rates (Lemma A.3). Thus, minimax theory based on effective support size extends the classical unrestricted theory while remaining informative for distributions whose effective support size is small.

Comparison with existing KL bounds. Several recent works derive finite-sample KL guarantees for uniform discrete diffusion. Cho & Wu (2026) establish KL rate of order $\widetilde{\mathcal{O}}(K^d/n)$, which they claim is nearly minimax optimal. However, their guarantees are established in the regime $n \gtrsim K^d$, which does not capture the regime $n \ll K^d$ typical of practical discrete diffusion. Wakasugi & Suzuki (2025) obtain a high-probability KL bound of the same order, together with a refined bound that removes the explicit polynomial dependence on K^d. This refinement requires decay of the target distribution's coefficients in an eigenbasis of the generator, a structural condition that is difficult to verify in practice. Moreover, the refined bound is not minimax optimal. Srikanth et al. (2026) derive a high-probability KL guarantee with explicit neural-network complexity, but require unrealistic full support assumption and incur polynomial dependence on the maximum initial score ratio, which may be as large as the ambient cardinality K^d. In contrast, our bound adapts to the effective support size $\mathfrak{s}_n(P_0)$ without such structural assumptions, are minimax optimal under effective support class $\mathcal{P}_{s_n}$, and yields rates $\mathcal{O}\left(n^{\alpha^{-1}-1} \log \frac{eK^d}{n^{1/\alpha}} \log n\right)$ and $\mathcal{O}\left(\frac{\log^2 n}{n} \log \frac{eK^d}{\log n}\right)$ for Zipfian and geometric decay, respectively, in the non-saturated regimes of (1).

4 Experiments

Controlled populations. We conduct a controlled experiment to evaluate whether uniform discrete diffusion models learn populations with smaller effective support sizes more accurately when the size of the ambient state space is held fixed. We first train a byte-level BPE tokenizer with vocabulary size $K = 2048$ on TinyStories (Eldan & Li, 2023). Using the resulting tokenized corpus, we train an 88M-parameter, 12-layer decoder-only Transformer as an autoregressive teacher model that predicts each token conditioned on the preceding tokens. We then use the frozen teacher at temperatures

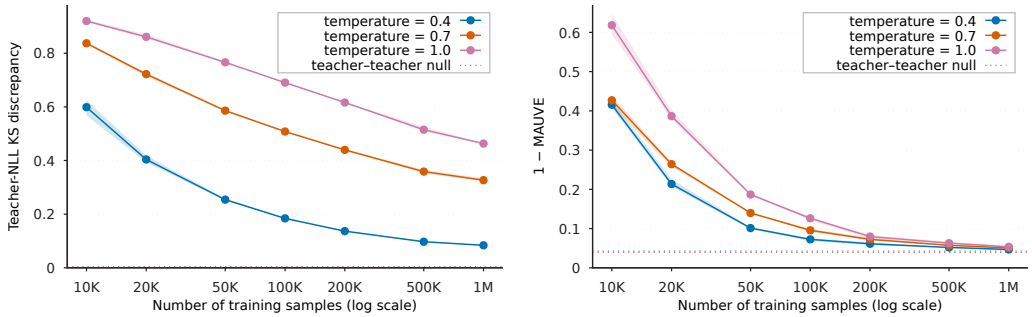


Figure 2: **Uniform discrete diffusion on controlled populations.** Teacher-NLL KS (left) and $1 - \mathrm{MAUVE}$ (right) versus training-set size for teacher temperatures $\{0.4, 0.7, 1.0\}$, with ambient state space 2048^{32}. Lines and markers show means over three runs after 50 training epochs, and shaded bands span the minimum and maximum across runs. Dotted lines indicate teacher–teacher null controls. Lower values indicate better agreement between population and sample distributions.

$\{0.4, 0.7, 1.0\}$ to define three population distributions over $d = 32$-token sequences, all on the same ambient state space of size $K^d = 2048^{32}$. Lower temperatures concentrate the teacher's next-token probabilities and therefore produce populations with smaller effective support sizes.

Training and Evaluation. For each population, we consider seven training-set sizes ($n = 10\mathrm{K}$–$1\mathrm{M}$). For every population and training-set size, we train a 96M-parameter uniform SEDD model (Lou et al., 2024) for 50 epochs. Each experimental setting is repeated three times. Directly computing full-sequence TV or KL is infeasible because it would require tractable access to the model probabilities over 2048^{32} possible sequences. We therefore draw 100K samples from each trained SEDD model and 100K fresh samples from the corresponding teacher population and compare them using two sample-based discrepancies. Teacher-NLL KS compares the distributions of sequence negative log-likelihood under the teacher; the corresponding population discrepancy lower-bounds, but does not estimate, full-sequence TV. The second metric, $1 - \mathrm{MAUVE}$, compares the samples in a fixed GPT-2-large representation (Pillutla et al., 2021). We apply both metrics to two independent sets of 100K samples from each teacher population as a teacher-versus-teacher null control, which indicates their finite-sample evaluation floors. Lower values indicate better agreement for both metrics.

Results. Figure 2 shows that both teacher-NLL KS and $1 - \mathrm{MAUVE}$ decrease as the training-set size increases. At every training-set size, uniform SEDD models trained on lower-temperature populations achieve smaller discrepancies under both metrics. Since these populations have smaller effective support sizes while sharing the same ambient state space, the results support effective support size as a meaningful measure of statistical difficulty for uniform discrete diffusion.

5 Conclusion

In this paper, we develop a statistical theory for uniform discrete diffusion models on finite product spaces whose ambient cardinality K^d can be astronomically large. Our analysis is governed by the effective support size $\mathfrak{s}_n(P_0)$, a sample-size-dependent measure of distributional complexity. We show that uniform discrete diffusion achieves the minimax-optimal TV rate $\mathcal{O}(\sqrt{\mathfrak{s}_n(P_0)/n})$. Under KL loss, early stopping at time $1/n$ yields the rate $\mathcal{O}\left(\frac{1}{n}\mathfrak{s}_n(P_0)\log(eK^d/\mathfrak{s}_n(P_0))\log n\right)$, which is within a factor of $\log n$ of the minimax benchmark. This KL guarantee captures generalization beyond the observed samples, since the empirical distribution generally assigns zero probability to unseen states and therefore has infinite KL risk. These minimax rates identify three regimes in which consistent distribution estimation is impossible, possible only under TV loss, or possible under both TV and KL losses. Our controlled experiment with uniform SEDD models complements these theoretical findings and supports effective support size as a meaningful measure of statistical difficulty beyond ambient cardinality alone. Important directions for future work include closing the remaining logarithmic gap under KL loss, controlling optimization error in score estimation, and extending the analysis to masking discrete diffusion models.

DISCLOSURE OF AI USE

We used generative AI tools to assist with writing and editing experimental code, checking mathematical proofs developed by the authors, retrieving potentially relevant related work, and polishing the manuscript for clarity and readability. The research ideas and original mathematical proofs were developed by the authors without generative AI assistance. All AI-assisted outputs were thoroughly reviewed and verified by the authors. We take full responsibility for the final content of this paper, including its text, mathematical claims and proofs, experimental results, citations, and accompanying code.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

Anna Ben-Hamou, Stéphane Boucheron, and Mesrob I. Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1):249–287, 2017. doi: 10.3150/15-BEJ743.

Dietrich Braess and Thomas Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

Dietrich Braess, Jürgen Forster, Tomas Sauer, and Hans U Simon. How to achieve minimax expected kullback-leibler distance from an unknown finite distribution. In *International Conference on Algorithmic Learning Theory*, pp. 380–394. Springer, 2002.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28266–28279, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b5b528767aa35f5b1a60fe0aaeca0563-Abstract-Conference.html.

Hongrui Chen and Lexing Ying. Convergence analysis of discrete diffusion model: Exact implementation through uniformization. *Journal of Machine Learning*, 4(2):108–127, 2025. doi: 10.4208/jml.240812.

Cholyeon Cho and Yuchen Wu. Minimax optimality of score-entropy discrete diffusion. *arXiv preprint arXiv:2608.20635*, 2026. URL https://arxiv.org/abs/2608.20635.

Daniil Dmitriev, Zhihan Huang, and Yuting Wei. Efficient sampling with discrete diffusion models: Sharp and adaptive guarantees. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 2038–2104, 2026. URL https://proceedings.mlr.press/v336/dmitriev26a.html.

Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*, 2023. URL https://arxiv.org/abs/2305.07759.

Moein Falahatgar, Mesrob I. Ohannessian, Alon Orlitsky, and Venkatadheeraj Pichapati. The power of absolute discounting: All-dimensional distribution estimation. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6660–6669, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/331316d4efb44682092a006307b9ae3a-Abstract.html.

Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976. doi: 10.1016/0021-9991(76)90041-3.

Daniel T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733, 2001. doi: 10.1063/1.1378322.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264, 1953. doi: 10.1093/biomet/40.3-4.237.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory*, 61(11):6343–6354, 2015. doi: 10.1109/TIT.2015.2478816.

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. DiffusionBERT: Improving generative masked language models with diffusion models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4521–4534, 2023. URL https://aclanthology.org/2023.acl-long.248/.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12454–12465, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/67d96d458abdef21792e6d8e590244e7-Abstract.html.

Arne Jensen. Markoff chains as an aid in the study of Markoff processes. *Scandinavian Actuarial Journal*, 1953(sup1):87–91, 1953. doi: 10.1080/03461238.1953.10419459.

Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1066–1100, 2015. URL https://proceedings.mlr.press/v40/Kamath15.html.

Samuel Karlin. Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, 17(4):373–401, 1967. URL https://www.jstor.org/stable/24902077.

Yuchen Liang, Yingbin Liang, Lifeng Lai, and Ness Shroff. Discrete diffusion models: Novel analysis and new sampler guarantees. In *Advances in Neural Information Processing Systems*, volume 38, pp. 183461–183498, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f1e7c90552850afcc2558d78950c519d-Abstract-Conference.html.

Sulin Liu, Juno Nam, Andrew Campbell, Hannes Stärk, Yilun Xu, Tommi Jaakkola, and Rafael Gómez-Bombarelli. Think while you generate: Discrete diffusion with planned denoising. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=MJNywBdSDy.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848, 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

Jaouad Mourtada. Estimation of discrete distributions in relative entropy, and the deviations of the missing mass. *Mathematical Statistics and Learning*, 9(1/2):69–144, 2026. doi: 10.4171/MSL/55.

J. R. Norris. *Markov Chains*, volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 1997. doi: 10.1017/CBO9780511810633.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26517–26582, 2023. URL https://proceedings.mlr.press/v202/oko23a.html.

Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is Good–Turing good. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2143–2151, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/d759175de8ea5b1d9a2660e45554894f-Abstract.html.

Liam Paninski. Variational minimax estimation of discrete distributions under KL loss. In *Advances in Neural Information Processing Systems*, volume 17, pp. 1033–1040, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/hash/c57168a952f5d46724cf35dfc3d48a7f-Abstract.html.

Le-Tuyet-Nhi Pham, Dario Shariatian, Antonio Ocello, Giovanni Conforti, and Alain Oliviero Durmus. Discrete Markov probabilistic models: An improved discrete score-based framework with sharp convergence bounds under minimal assumptions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49195–49258, 2025. URL https://proceedings.mlr.press/v267/pham25a.html.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4816–4828, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/260c2432a0eecc28ce03c10dadc078a4-Abstract.html.

Yinuo Ren, Haoxuan Chen, Grant M. Rotskoff, and Lexing Ying. How discrete and continuous diffusion meet: Comprehensive analysis of discrete diffusion models via a stochastic integral framework. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6awxwQEI82.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Aadithya Srikanth, Mudit Gaur, and Vaneet Aggarwal. Discrete state diffusion models: A sample complexity perspective. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3547–3555, 2026. URL https://proceedings.mlr.press/v300/srikanth26a.html.

Zhiqing Sun and Yiming Yang. DIFUSCO: Graph-based diffusion solvers for combinatorial optimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 3706–3731, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0ba520d93c3df592c83a611961314c98-Abstract-Conference.html.

Dirk van der Hoeven, Julia Olkhovskaia, and Tim van Erven. Nearly minimax discrete distribution estimation in kullback-leibler divergence with high probability. *arXiv preprint arXiv:2507.17316*, 2025.

Clément Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. DiGress: Discrete denoising diffusion for graph generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UaAD-Nu86WX.

Shintaro Wakasugi and Taiji Suzuki. State size independent statistical error bound for discrete diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 154062–154097, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/caabce23491e5418f4c9ea348359fcd2-Abstract-Conference.html.

Jiayuan Ye, Vitaly Feldman, and Kunal Talwar. Instance-optimality for private kl distribution estimation. *arXiv preprint arXiv:2505.23620*, 2025.

Zikun Zhang, Zixiang Chen, and Quanquan Gu. Convergence of score-based discrete diffusion models: A discrete-time analysis. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pq1WUegkza.

Zixuan Zhang, Hengyu Fu, Zhuoran Yang, Mengdi Wang, Tuo Zhao, and Minshuo Chen. Generalization bounds for discrete diffusion: Statistical advantage of masking. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Ofhq7nBVHu.

APPENDICES

A Preliminary Results **13**

A.1 Some Facts on Effective Support Size 13

A.2 Properties of Uniform Forward Process 15

A.3 Auxiliary Results for KL Bounds 16

A.4 Score-Entropy Identities and Sampler Guarantees 19

A.5 ReLU Network Constructions 20

B Proofs for TV Results **22**

B.1 Upper Bound: Proofs of Theorem 3 and Corollary 4 22

B.2 Minimax Lower Bound: Proof of Theorem 5 23

C Proofs for KL Results **24**

C.1 Upper Bound: Proofs of Theorem 6 and Corollary 7 24

C.2 Minimax Lower Bound: Proof of Theorem 8 26

D Exact One-Step Empirical Reverse Sampling **28**

D.1 Exact reverse kernel and sampling algorithm 28

D.2 Statistical guarantees 29

Notation. We retain the notation of the main text and introduce only appendix-specific shorthand here. For any probability distribution μ on $[K]^d$, write $\mu_t := \mu \boldsymbol{P}_t$ for its marginal after evolving under the forward process for time t. When the data distribution is P_0, we abbreviate its point masses as $p_{\boldsymbol{x}} := P_0(\boldsymbol{x})$. For $t > 0$, define

$$B_0(t) := \left(\frac{K}{1-e^{-t}}\right)^d, \qquad B_1(t) := \frac{1+(K-1)e^{-t}}{1-e^{-t}}. \tag{.4}$$

For $u \in \mathbb{R}$, write $[u]_+ := \max\{u, 0\}$; in the neural-network constructions, $\rho(u) := \max\{u, 0\}$ denotes the ReLU activation.

A PRELIMINARY RESULTS

A.1 SOME FACTS ON EFFECTIVE SUPPORT SIZE

Lemma A.1 (Effective support size under representative decay regimes). *Let* $p_{(1)} \geq \cdots \geq p_{(|\operatorname{supp}(P_0)|)} > 0$ *denote nonzero masses of* P_0 *in decreasing order. Suppose* $n \geq 2$. *In each*

condition below, $\asymp$ means that the comparison holds uniformly over $1 \le j \le |\operatorname{supp}(P_0)|$ with universal constants. Then

$$\mathfrak{s}_n(P_0) \asymp \begin{cases} \min\{|\operatorname{supp}(P_0)|, n\}, & \text{if } p_{(j)} \asymp |\operatorname{supp}(P_0)|^{-1}, \\ \min\{|\operatorname{supp}(P_0)|, n^{1/\alpha}\}, & \text{if } p_{(j)} \asymp j^{-\alpha}, \\ \min\{|\operatorname{supp}(P_0)|, \log n\}, & \text{if } p_{(j)} \asymp e^{-cj}, \end{cases}$$

where $\alpha > 1$ and $c > 0$ are fixed. The implicit constants in the conclusion depend only on the comparison constants and, in the second and third cases, on α and c, respectively.

Proof. Since $u \asymp v$ implies $\min\{nu, 1\} \asymp \min\{nv, 1\}$, we may work directly with the representative sequences in the three cases.

If $p_{(j)} \asymp |\operatorname{supp}(P_0)|^{-1}$, direct calculation gives $\mathfrak{s}_n(P_0) \asymp \min\{|\operatorname{supp}(P_0)|, n\}$.

If $p_{(j)} \asymp j^{-\alpha}$, then

$$\mathfrak{s}_n(P_0) \asymp \sum_{j=1}^{|\operatorname{supp}(P_0)|} \min\{nj^{-\alpha}, 1\}.$$

If $|\operatorname{supp}(P_0)| \le n^{1/\alpha}$, all summands equal 1, so $\mathfrak{s}_n(P_0) \asymp |\operatorname{supp}(P_0)|$. If $|\operatorname{supp}(P_0)| > n^{1/\alpha}$, the first $\lfloor n^{1/\alpha} \rfloor$ summands equal 1, while

$$n \sum_{j > \lfloor n^{1/\alpha} \rfloor} j^{-\alpha} \lesssim n^{1/\alpha}.$$

Thus $\mathfrak{s}_n(P_0) \asymp n^{1/\alpha}$, and hence $\mathfrak{s}_n(P_0) \asymp \min\{|\operatorname{supp}(P_0)|, n^{1/\alpha}\}$.

If $p_{(j)} \asymp e^{-cj}$, then

$$\mathfrak{s}_n(P_0) \asymp \sum_{j=1}^{|\operatorname{supp}(P_0)|} \min\{ne^{-cj}, 1\}.$$

If $|\operatorname{supp}(P_0)| \le c^{-1} \log n$, all summands equal 1, so $\mathfrak{s}_n(P_0) \asymp |\operatorname{supp}(P_0)|$. If $|\operatorname{supp}(P_0)| > c^{-1} \log n$, the summands equal 1 for $j \le \lfloor c^{-1} \log n \rfloor$, while

$$n \sum_{j > \lfloor c^{-1} \log n \rfloor} e^{-cj} \lesssim 1.$$

Thus $\mathfrak{s}_n(P_0) \asymp \log n$, and hence $\mathfrak{s}_n(P_0) \asymp \min\{|\operatorname{supp}(P_0)|, \log n\}$. □

Lemma A.2 (Effective support size and distinct observations)**.** *Let $X_1, \dots, X_n$ be i.i.d. samples from P_0, and let $D_n := |\{X_1, \dots, X_n\}|$ denote the number of distinct observations. Then*

$$(1 - e^{-1})\mathfrak{s}_n(P_0) \le \mathbb{E}_{P_0} D_n \le \mathfrak{s}_n(P_0).$$

Moreover, for every $\eta \in (0, 1)$, with probability at least $1 - \eta$,

$$D_n \le 2\big(\mathfrak{s}_n(P_0) + \log \eta^{-1}\big).$$

Proof. We have $\mathbb{E}_{P_0} D_n = \sum_x \{1 - (1 - P_0(x))^n\}$. Since $1 - (1-u)^n \le \min\{nu, 1\}$ for $u \in [0, 1]$, it follows that $\mathbb{E}_{P_0} D_n \le \mathfrak{s}_n(P_0)$. Conversely, $1 - (1-u)^n \ge 1 - e^{-nu} \ge (1 - e^{-1}) \min\{nu, 1\}$, so $\mathbb{E}_{P_0} D_n \ge (1 - e^{-1})\mathfrak{s}_n(P_0)$. Finally, Mourtada (2026, Lemma 19) gives, with probability at least $1 - \eta$,

$$D_n \le 2\big(\mathbb{E}_{P_0} D_n + \log \eta^{-1}\big),$$

and the claim follows from $\mathbb{E}_{P_0} D_n \le \mathfrak{s}_n(P_0)$. □

Lemma A.3 (Recovery of the unrestricted KL rate)**.** *Let $s_{\max} := \min\{K^d, n\}$. Then,*

$$\log\left(1 + \frac{K^d}{n}\right) \le \frac{s_{\max}}{n} \log\left(\frac{eK^d}{s_{\max}}\right) \le 2 \log\left(1 + \frac{K^d}{n}\right).$$

Proof. If $K^d \le n$, then $s_{\max} = K^d$ and

$$\frac{s_{\max}}{n} \log\left(\frac{eK^d}{s_{\max}}\right) = \frac{K^d}{n}.$$

Since $K^d/n \le 1$, the inequalities $x/2 \le \log(1+x) \le x$ for $x \in [0,1]$ give the result.

If $K^d \ge n$, then $s_{\max} = n$ and

$$\frac{s_{\max}}{n} \log\left(\frac{eK^d}{s_{\max}}\right) = 1 + \log\left(\frac{K^d}{n}\right).$$

Writing $x := K^d/n \ge 1$, we have $1 + x \le ex \le (1+x)^2$, and hence

$$\log(1+x) \le 1 + \log x \le 2\log(1+x).$$

Combining the two cases proves the claim. □

A.2 Properties of Uniform Forward Process

Lemma A.4. *For every $t > 0$ and $\boldsymbol{x} \in [K]^d$,*

$$P_t(\boldsymbol{x}) \ge B_0(t)^{-1}.$$

Moreover, for every $t > 0$ and $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in [K]^d$ with $d_{\mathrm{Ham}}(\boldsymbol{x}, \boldsymbol{y}) = 1$,

$$B_1(t)^{-1} \le \frac{P_t(\boldsymbol{y})}{P_t(\boldsymbol{x})} \le B_1(t), \qquad B_1(t)^{-1} \le \frac{\boldsymbol{P}_t(\boldsymbol{z}, \boldsymbol{y})}{\boldsymbol{P}_t(\boldsymbol{z}, \boldsymbol{x})} \le B_1(t).$$

Proof. For one coordinate, $\boldsymbol{P}_t^{\mathrm{coord}}(a,b) = e^{-t}\mathbf{1}\{a=b\} + (1-e^{-t})/K \ge (1-e^{-t})/K$. Thus, by coordinatewise independence, $\boldsymbol{P}_t(\boldsymbol{z}, \boldsymbol{x}) \ge ((1-e^{-t})/K)^d = B_0(t)^{-1}$, and averaging over $\boldsymbol{z} \sim P_0$ gives $P_t(\boldsymbol{x}) \ge B_0(t)^{-1}$.

Now suppose $d_{\mathrm{Ham}}(\boldsymbol{x}, \boldsymbol{y}) = 1$, and let j be the unique coordinate on which they differ. All other coordinate factors cancel, so

$$\frac{\boldsymbol{P}_t(\boldsymbol{z}, \boldsymbol{y})}{\boldsymbol{P}_t(\boldsymbol{z}, \boldsymbol{x})} = \frac{\boldsymbol{P}_t^{\mathrm{coord}}(z^j, y^j)}{\boldsymbol{P}_t^{\mathrm{coord}}(z^j, x^j)}.$$

Since the smallest and largest entries of $\boldsymbol{P}_t^{\mathrm{coord}}$ are $(1-e^{-t})/K$ and $e^{-t} + (1-e^{-t})/K$, respectively, the last ratio lies in $[B_1(t)^{-1}, B_1(t)]$.

Finally,

$$\frac{P_t(\boldsymbol{y})}{P_t(\boldsymbol{x})} = \sum_{\boldsymbol{z} \in [K]^d} \frac{P_0(\boldsymbol{z})\boldsymbol{P}_t(\boldsymbol{z}, \boldsymbol{x})}{P_t(\boldsymbol{x})} \frac{\boldsymbol{P}_t(\boldsymbol{z}, \boldsymbol{y})}{\boldsymbol{P}_t(\boldsymbol{z}, \boldsymbol{x})}.$$

The coefficients form a probability distribution over $\boldsymbol{z}$, so the same bounds hold for $P_t(\boldsymbol{y})/P_t(\boldsymbol{x})$. □

Lemma A.5 (Uniform mixing of the forward process)**.** *For every $t \ge 0$,*

$$d_{\mathrm{TV}}(P_t, \pi) \le 1 - (1-e^{-t})^d \le de^{-t}.$$

Proof. For one coordinate, $\boldsymbol{P}_t^{\mathrm{coord}}(a, \cdot) = e^{-t}\delta_a + (1-e^{-t})\,\mathrm{Unif}([K])$. Hence, with probability $(1-e^{-t})^d$, all d coordinates are refreshed independently from the uniform distribution, regardless of the initial state. Therefore, for some distribution Q_t,

$$P_t = (1-e^{-t})^d \pi + \left[1 - (1-e^{-t})^d\right] Q_t.$$

It follows that $d_{\mathrm{TV}}(P_t, \pi) \le 1 - (1-e^{-t})^d \le de^{-t}$, where the last inequality follows from $1 - (1-u)^d \le du$. □

Lemma A.6. *The empirical distribution $\widehat{P}_0$ satisfies*

$$\mathbb{E}_{P_0}\left[d_{\mathrm{TV}}(P_0, \widehat{P}_0)\right] \le \frac{3}{2}\sqrt{\frac{\mathfrak{s}_n(P_0)}{n}}.$$

Proof. Let $N_{\boldsymbol{x}} := \sum_{i=1}^{n} \mathbf{1}\{X_0^{(i)} = \boldsymbol{x}\}$ and $H := \{\boldsymbol{x} \in [K]^d : nP_0(\boldsymbol{x}) \geq 1\}$. For every $\boldsymbol{x} \in [K]^d$, the triangle inequality and Jensen's inequality give

$$\mathbb{E}_{P_0}\left|\frac{N_{\boldsymbol{x}}}{n} - P_0(\boldsymbol{x})\right| \leq \min\left\{2P_0(\boldsymbol{x}), \sqrt{\frac{P_0(\boldsymbol{x})}{n}}\right\}.$$

Therefore,

$$\mathbb{E}_{P_0}\left[d_{\mathrm{TV}}(P_0, \widehat{P}_0)\right] \leq \sum_{\boldsymbol{x}\notin H} P_0(\boldsymbol{x}) + \frac{1}{2}\sum_{\boldsymbol{x}\in H}\sqrt{\frac{P_0(\boldsymbol{x})}{n}}.$$

The first term is at most $\mathfrak{s}_n(P_0)/n \leq \sqrt{\mathfrak{s}_n(P_0)/n}$. Since $|H| \leq \mathfrak{s}_n(P_0)$, the Cauchy–Schwarz inequality gives

$$\sum_{\boldsymbol{x}\in H}\sqrt{\frac{P_0(\boldsymbol{x})}{n}} \leq \sqrt{\frac{|H|}{n}\sum_{\boldsymbol{x}\in H} P_0(\boldsymbol{x})} \leq \sqrt{\frac{\mathfrak{s}_n(P_0)}{n}}.$$

Combining the two bounds proves the claim. □

A.3 Auxiliary Results for KL Bounds

Lemmas A.7 to A.10 are used for proving Lemma A.11, which gives the KL divergence upper bound for the empirical distribution at time $1/n$, $\widehat{P}_{1/n}$.

Lemma A.7 (Chernoff bounds for a binomial random variable)**.** *Let N be a binomial random variable. For every $\delta \in (0, 1)$,*

$$\mathbb{P}\{N \leq (1-\delta)\mathbb{E}N\} \leq \exp\left(-\frac{\delta^2\mathbb{E}N}{2}\right), \qquad \mathbb{P}\{N \geq (1+\delta)\mathbb{E}N\} \leq \exp\left(-\frac{\delta^2\mathbb{E}N}{2+\delta}\right).$$

Proof. Write $N \sim \mathrm{Bin}(M, p)$, so $\mathbb{E}N = Mp$. For every $\lambda \in \mathbb{R}$, $\mathbb{E}e^{\lambda N} = (1 - p + pe^{\lambda})^M \leq \exp\{(\mathbb{E}N)(e^{\lambda} - 1)\}$. Applying Markov's inequality with $\lambda = \log(1+\delta)$ and using $(1+\delta)\log(1+\delta) - \delta \geq \delta^2/(2+\delta)$ gives the upper-tail bound. Applying the same argument to $e^{-\lambda N}$ with $\lambda = -\log(1-\delta)$ and using $\delta + (1-\delta)\log(1-\delta) \geq \delta^2/2$ gives the lower-tail bound. □

Lemma A.8. *Let N be a binomial random variable. If $\mathbb{E}N \geq 1$,*

$$\mathbb{E}\left[\mathbf{1}\{N > 0\}\log\frac{\mathbb{E}N}{N}\right] \lesssim \frac{1}{\mathbb{E}N}.$$

If $0 < \mathbb{E}N < 1$, then $\mathbf{1}\{N > 0\}\log(\mathbb{E}N/N) \leq 0$.

Proof. If $0 < \mathbb{E}N < 1$, then $N > 0$ implies $N \geq 1 > \mathbb{E}N$, proving the second claim. Suppose $\lambda := \mathbb{E}N \geq 1$, and let $A := \{\lambda/2 \leq N \leq 3\lambda/2\}$ and $Z := (N - \lambda)/\lambda$. On A, since $|Z| \leq 1/2$ and $-\log(1+Z) \leq -Z + Z^2$,

$$\mathbb{E}\left[\mathbf{1}_A \log\frac{\lambda}{N}\right] \leq -\mathbb{E}[\mathbf{1}_A Z] + \mathbb{E}[Z^2].$$

Since $\mathbb{E}Z = 0$, $\mathbb{E}Z^2 \leq 1/\lambda$, and Lemma A.7 gives $\mathbb{P}(A^c) \leq 2e^{-\lambda/10}$, the Cauchy–Schwarz inequality yields

$$-\mathbb{E}[\mathbf{1}_A Z] = \mathbb{E}[\mathbf{1}_{A^c} Z] \leq \mathbb{E}[\mathbf{1}_{A^c}|Z|] \leq \sqrt{\frac{2}{\lambda}}e^{-\lambda/20} \lesssim \frac{1}{\lambda}.$$

Hence $\mathbb{E}[\mathbf{1}_A \log(\lambda/N)] \lesssim 1/\lambda$.

On A^c, the logarithm is positive only when $0 < N < \lambda/2$, where $\log(\lambda/N) \leq \log\lambda$. Therefore, Lemma A.7 gives

$$\mathbb{E}\left[\mathbf{1}_{A^c}\mathbf{1}\{N > 0\}\log\frac{\lambda}{N}\right] \leq \log\lambda\, e^{-\lambda/8} \lesssim \frac{1}{\lambda}.$$

Combining the two bounds proves the claim. □

Lemma A.9 (Weighted Caro–Wei bound). *Let $G = (V, E)$ be a finite simple graph with positive vertex weights $(w_v)_{v\in V}$, and let $N[v] := \{v\} \cup \{u \in V : \{u, v\} \in E\}$ denote the closed neighborhood of v. Let*

$$\alpha(G) := \max\{|I| : I \subseteq V,\ \{u, v\} \notin E \text{ for all distinct } u, v \in I\}$$

denote the independence number of G. Then

$$\sum_{v\in V} \frac{w_v}{\sum_{u\in N[v]} w_u} \le \alpha(G).$$

Proof. Let $(T_v)_{v\in V}$ be independent exponential random variables with rates $(w_v)_{v\in V}$, and define $I := \{v \in V : T_v = \min_{u\in N[v]} T_u\}$. Since the T_v's are continuous, the minima are unique almost surely. If $\{u, v\} \in E$ and $u, v \in I$, then $T_u < T_v$ and $T_v < T_u$, a contradiction, so I is independent almost surely. By the exponential race identity,

$$\mathbb{P}(v \in I) = \frac{w_v}{\sum_{u\in N[v]} w_u}.$$

Therefore,

$$\sum_{v\in V} \frac{w_v}{\sum_{u\in N[v]} w_u} = \mathbb{E}|I| \le \alpha(G).$$

□

Lemma A.10. *Let $X_0^{(1)}, \dots, X_0^{(n)}, Y$ be i.i.d. from P_0, and define*

$$R(Y) := \min_{1\le i\le n} d_{\mathrm{Ham}}\left(Y, X_0^{(i)}\right).$$

Then

$$\mathbb{E}_{P_0} R(Y) \lesssim \frac{\mathfrak{s}_n(P_0)}{n}\left\{1 + \frac{\log\big(eK^d/\mathfrak{s}_n(P_0)\big)}{\log K}\right\}.$$

Proof. For $r = 0, \dots, d$, let $B_r(\boldsymbol{y}) := \{\boldsymbol{z} \in [K]^d : d_{\mathrm{Ham}}(\boldsymbol{z}, \boldsymbol{y}) \le r\}$. For $r = 0, \dots, d-1$, independence gives

$$\mathbb{P}_{P_0}\{R(Y) > r\} \le \sum_{\boldsymbol{y}} P_0(\boldsymbol{y}) e^{-nP_0(B_r(\boldsymbol{y}))}.$$

Since $P_0(B_r(\boldsymbol{y})) \ge P_0(\boldsymbol{y})$ and $ue^{-u} \le \min\{u, 1\}$,

$$\sum_{\boldsymbol{y}} P_0(\boldsymbol{y}) e^{-nP_0(B_r(\boldsymbol{y}))} \le \frac{\mathfrak{s}_n(P_0)}{n}.$$

For a second bound, let G_r be the graph on $\mathrm{supp}(P_0)$ in which two distinct states are adjacent if their Hamming distance is at most r. Since the closed neighborhood of $\boldsymbol{y}$ has total weight $P_0(B_r(\boldsymbol{y}))$, Lemma A.9 and $ue^{-u} \le 1$ give

$$\sum_{\boldsymbol{y}} P_0(\boldsymbol{y}) e^{-nP_0(B_r(\boldsymbol{y}))} \le \frac{\alpha(G_r)}{n}.$$

If A is independent in G_r, the balls $B_{\lfloor r/2\rfloor}(\boldsymbol{x})$, $\boldsymbol{x} \in A$, are pairwise disjoint, and each contains at least $K^{\lfloor r/2\rfloor}$ states. Hence $\alpha(G_r) \le K^{d-\lfloor r/2\rfloor}$, and therefore

$$\mathbb{P}_{P_0}\{R(Y) > r\} \le \frac{1}{n}\min\left\{\mathfrak{s}_n(P_0), K^{d-\lfloor r/2\rfloor}\right\}.$$

Summing the tail probabilities yields

$$\mathbb{E}_{P_0} R(Y) \le \frac{2}{n}\sum_{\ell=0}^{d} \min\{\mathfrak{s}_n(P_0), K^{d-\ell}\}.$$

There are at most $1 + \log(K^d/\mathfrak{s}_n(P_0))/\log K$ indices for which $K^{d-\ell} \geq \mathfrak{s}_n(P_0)$, and their total contribution is at most

$$\mathfrak{s}_n(P_0)\left\{1 + \frac{\log(K^d/\mathfrak{s}_n(P_0))}{\log K}\right\}.$$

The remaining terms form a geometric series with largest term smaller than $\mathfrak{s}_n(P_0)$ and ratio $K^{-1} \leq 1/2$, so their sum is at most $2\mathfrak{s}_n(P_0)$. Therefore,

$$\sum_{\ell=0}^{d} \min\{\mathfrak{s}_n(P_0), K^{d-\ell}\} \lesssim \mathfrak{s}_n(P_0)\left\{1 + \frac{\log\big(eK^d/\mathfrak{s}_n(P_0)\big)}{\log K}\right\},$$

which proves the claim. □

Lemma A.11. *The empirical distribution at time $1/n$, $\widehat{P}_{1/n}$, satisfies*

$$\mathbb{E}_{P_0} D_{\mathrm{KL}}\left(P_0 \| \widehat{P}_{1/n}\right) \lesssim \frac{\mathfrak{s}_n(P_0)\log\big(eK^d/\mathfrak{s}_n(P_0)\big)\log n}{n}.$$

Proof. For $\boldsymbol{y} \in [K]^d$, let $N_{\boldsymbol{y}} := \sum_{i=1}^n \mathbf{1}\{X_0^{(i)} = \boldsymbol{y}\}$ and $R_{\boldsymbol{y}} := \min_{1\leq i\leq n} d_{\mathrm{Ham}}(\boldsymbol{y}, X_0^{(i)})$. By coordinatewise independence, choosing an observation nearest to $\boldsymbol{y}$, and using the $N_{\boldsymbol{y}}$ observations equal to $\boldsymbol{y}$ when $N_{\boldsymbol{y}} > 0$, give

$$\widehat{P}_{1/n}(\boldsymbol{y}) \geq \frac{e^{-d/n}}{n}\left(\frac{1-e^{-1/n}}{K}\right)^{R_{\boldsymbol{y}}}, \qquad N_{\boldsymbol{y}} > 0 \implies \widehat{P}_{1/n}(\boldsymbol{y}) \geq \frac{N_{\boldsymbol{y}}}{n}e^{-d/n}.$$

We decompose the KL divergence according to whether $\boldsymbol{y}$ appears in the sample. For the unobserved states, the first bound and $1 - e^{-1/n} \geq 1/(2n)$ give

$$\sum_{\boldsymbol{y}:N_{\boldsymbol{y}}=0} P_0(\boldsymbol{y}) \log \frac{P_0(\boldsymbol{y})}{\widehat{P}_{1/n}(\boldsymbol{y})} \leq \sum_{\boldsymbol{y}:N_{\boldsymbol{y}}=0} P_0(\boldsymbol{y})[\log(nP_0(\boldsymbol{y}))]_+ + \frac{d}{n} + \log(2Kn) \sum_{\boldsymbol{y}:N_{\boldsymbol{y}}=0} P_0(\boldsymbol{y}) R_{\boldsymbol{y}}.$$

Since $\mathbb{P}_{P_0}\{N_{\boldsymbol{y}} = 0\} = (1 - P_0(\boldsymbol{y}))^n$,

$$\mathbb{E}_{P_0} \sum_{\boldsymbol{y}:N_{\boldsymbol{y}}=0} P_0(\boldsymbol{y})[\log(nP_0(\boldsymbol{y}))]_+ \leq \frac{1}{n} \sum_{\boldsymbol{y}:nP_0(\boldsymbol{y})\geq 1} nP_0(\boldsymbol{y}) e^{-nP_0(\boldsymbol{y})} \log(nP_0(\boldsymbol{y})) \lesssim \frac{\mathfrak{s}_n(P_0)}{n}.$$

Moreover, Lemma A.10 gives

$$\mathbb{E}_{P_0} \sum_{\boldsymbol{y}} P_0(\boldsymbol{y}) R_{\boldsymbol{y}} \lesssim \frac{\mathfrak{s}_n(P_0)}{n}\left\{1 + \frac{\log\big(eK^d/\mathfrak{s}_n(P_0)\big)}{\log K}\right\}.$$

For the observed states, the second lower bound gives

$$\sum_{\boldsymbol{y}:N_{\boldsymbol{y}}>0} P_0(\boldsymbol{y}) \log \frac{P_0(\boldsymbol{y})}{\widehat{P}_{1/n}(\boldsymbol{y})} \leq \frac{d}{n} + \sum_{\boldsymbol{y}:N_{\boldsymbol{y}}>0} P_0(\boldsymbol{y}) \log \frac{nP_0(\boldsymbol{y})}{N_{\boldsymbol{y}}}.$$

If $nP_0(\boldsymbol{y}) < 1$, the corresponding summand is nonpositive. If $nP_0(\boldsymbol{y}) \geq 1$, Lemma A.8 gives its expectation at most C/n. Since at most $\mathfrak{s}_n(P_0)$ states satisfy $nP_0(\boldsymbol{y}) \geq 1$, the expected observed contribution is $\lesssim (d + \mathfrak{s}_n(P_0))/n$.

Combining the two parts,

$$\mathbb{E}_{P_0} D_{\mathrm{KL}}(P_0 \| \widehat{P}_{1/n}) \lesssim \frac{d + \mathfrak{s}_n(P_0)}{n} + \frac{\mathfrak{s}_n(P_0)\log(Kn)}{n}\left\{1 + \frac{\log\big(eK^d/\mathfrak{s}_n(P_0)\big)}{\log K}\right\}.$$

Since $u \mapsto u\log(eK^d/u)$ is increasing on $[1, K^d]$, we have $d \lesssim \mathfrak{s}_n(P_0)\log(eK^d/\mathfrak{s}_n(P_0))$. Also, using $\mathfrak{s}_n(P_0) \leq n$ and $K, n \geq 2$,

$$\log(Kn)\left\{1 + \frac{\log\big(eK^d/\mathfrak{s}_n(P_0)\big)}{\log K}\right\} \lesssim \log\left(\frac{eK^d}{\mathfrak{s}_n(P_0)}\right)\log n.$$

The claimed bound follows. □

Since KL divergence does not satisfy a triangle inequality, we use Lemma A.13 instead. Its proof relies on the entropy variational inequality in Lemma A.12.

Lemma A.12 (Entropy variational inequality)**.** *Let μ and ν be probability distributions on a finite state space. Then, for every real-valued function f,*

$$\mathbb{E}_\mu f \le D_{\mathrm{KL}}(\mu\|\nu) + \log \mathbb{E}_\nu e^f.$$

Proof. The claim is immediate if $D_{\mathrm{KL}}(\mu\|\nu) = \infty$. Otherwise, define $\nu_f(\boldsymbol{x}) := \nu(\boldsymbol{x})e^{f(\boldsymbol{x})}/\mathbb{E}_\nu e^f$. Then

$$0 \le D_{\mathrm{KL}}(\mu\|\nu_f) = D_{\mathrm{KL}}(\mu\|\nu) - \mathbb{E}_\mu f + \log \mathbb{E}_\nu e^f,$$

which proves the claim. □

Lemma A.13. *Let μ, ν, ξ be probability distributions on a finite state space. Suppose that, for some $B \ge 1$ and $\eta < \infty$,*

$$\nu(\boldsymbol{x}) \ge B^{-1} \quad \textit{for every } \boldsymbol{x}, \qquad D_{\mathrm{KL}}(\nu\|\xi) \le \frac{\eta}{B}.$$

Then there exists $C_\eta < \infty$, depending only on η, such that

$$D_{\mathrm{KL}}(\mu\|\xi) \le 2D_{\mathrm{KL}}(\mu\|\nu) + C_\eta D_{\mathrm{KL}}(\nu\|\xi).$$

Proof. Let $h(u) := u - 1 - \log u$ and $u_{\boldsymbol{x}} := \xi(\boldsymbol{x})/\nu(\boldsymbol{x})$. Since ν has full support and $D_{\mathrm{KL}}(\nu\|\xi) < \infty$, we also have $\xi(\boldsymbol{x}) > 0$ for every $\boldsymbol{x}$. Moreover,

$$h(u_{\boldsymbol{x}}) = \frac{D_{\mathrm{Br}}(\nu(\boldsymbol{x}), \xi(\boldsymbol{x}))}{\nu(\boldsymbol{x})} \le B D_{\mathrm{KL}}(\nu\|\xi) \le \eta.$$

The set $\{u > 0 : h(u) \le \eta\}$ is compact, and $u \mapsto (1-u)^2/(uh(u))$ extends continuously to $u = 1$ with value 2. Hence this function is bounded on the set by a constant depending only on η, and therefore

$$\chi^2(\nu, \xi) := \sum_{\boldsymbol{x}} \frac{(\nu(\boldsymbol{x}) - \xi(\boldsymbol{x}))^2}{\xi(\boldsymbol{x})} \le C_\eta D_{\mathrm{KL}}(\nu\|\xi).$$

Applying Lemma A.12 with $f(\boldsymbol{x}) := \log(\nu(\boldsymbol{x})/\xi(\boldsymbol{x}))$ gives

$$D_{\mathrm{KL}}(\mu\|\xi) \le 2D_{\mathrm{KL}}(\mu\|\nu) + \log\bigl(1 + \chi^2(\nu, \xi)\bigr) \le 2D_{\mathrm{KL}}(\mu\|\nu) + C_\eta D_{\mathrm{KL}}(\nu\|\xi).$$

□

A.4 Score-Entropy Identities and Sampler Guarantees

The following lemma is the discrete diffusion score-entropy loss analogue of the standard denoising identity.

Lemma A.14. *Let μ be a distribution on $[K]^d$, and define $\mu_t := \mu \boldsymbol{P}_t$ and $\sigma_t^{\star,\mu}(\boldsymbol{y}, \boldsymbol{x}) := \mu_t(\boldsymbol{y})/\mu_t(\boldsymbol{x})$. For every $t > 0$, there exists a finite quantity $C_t(\mu)$, independent of the candidate score σ, such that*

$$\mathcal{L}_t^{\mathrm{DSE}}(\sigma; \mu) = \mathcal{L}_t^{\mathrm{SE}}(\sigma; \mu) + C_t(\mu).$$

Consequently, for any positive score functions σ and $\widetilde{\sigma}$,

$$\mathcal{L}_{\mathcal{T}}^{\mathrm{DSE}}(\sigma; \mu) - \mathcal{L}_{\mathcal{T}}^{\mathrm{DSE}}(\widetilde{\sigma}; \mu) = \mathcal{L}_{\mathcal{T}}^{\mathrm{SE}}(\sigma; \mu) - \mathcal{L}_{\mathcal{T}}^{\mathrm{SE}}(\widetilde{\sigma}; \mu).$$

Proof. Fix $t > 0$, let $X_0 \sim \mu$, and conditionally on X_0, let $X_t \sim \boldsymbol{P}_t(X_0, \cdot)$. For fixed $\boldsymbol{x}, \boldsymbol{y}$, the conditional score satisfies

$$\sigma_{t|0}^\star(\boldsymbol{y}, \boldsymbol{x} \mid X_0) = \frac{\boldsymbol{P}_t(X_0, \boldsymbol{y})}{\boldsymbol{P}_t(X_0, \boldsymbol{x})}.$$

Conditioning on $X_t = \boldsymbol{x}$ and using Bayes' rule,

$$\mathbb{P}(X_0 = \boldsymbol{z} \mid X_t = \boldsymbol{x}) = \frac{\mu(\boldsymbol{z})\boldsymbol{P}_t(\boldsymbol{z}, \boldsymbol{x})}{\mu_t(\boldsymbol{x})}.$$

Therefore,

$$\mathbb{E}\Big[\sigma^\star_{t|0}(\boldsymbol{y},\boldsymbol{x} \mid X_0) \mid X_t = \boldsymbol{x}\Big] = \sum_{\boldsymbol{z}} \frac{\boldsymbol{P}_t(\boldsymbol{z},\boldsymbol{y})}{\boldsymbol{P}_t(\boldsymbol{z},\boldsymbol{x})} \frac{\mu(\boldsymbol{z})\boldsymbol{P}_t(\boldsymbol{z},\boldsymbol{x})}{\mu_t(\boldsymbol{x})}$$
$$= \frac{\mu_t(\boldsymbol{y})}{\mu_t(\boldsymbol{x})} = \sigma_t^{\star,\mu}(\boldsymbol{y},\boldsymbol{x}).$$

Conditioning the DSE loss on X_t and expanding the Bregman divergence in the SE loss therefore gives

$$\mathcal{L}_t^{\mathrm{DSE}}(\sigma;\mu) = \mathcal{L}_t^{\mathrm{SE}}(\sigma;\mu) + C_t(\mu),$$

where

$$C_t(\mu) := \mathbb{E}_{X_t\sim\mu_t} \sum_{\boldsymbol{y}:\boldsymbol{Q}(\boldsymbol{y},X_t)>0} \boldsymbol{Q}(\boldsymbol{y},X_t)\sigma_t^{\star,\mu}(\boldsymbol{y},X_t)\big(1-\log\sigma_t^{\star,\mu}(\boldsymbol{y},X_t)\big).$$

Since $t>0$, all forward probabilities are positive, so $C_t(\mu)$ is finite. Applying the identity to both σ and $\widetilde{\sigma}$, subtracting to cancel $C_t(\mu)$, and summing over the grid with weights $t_k - t_{k-1}$ proves the second statement. □

We next record Dmitriev et al. (2026, Theorem 1) in our notation. As noted there, the same guarantee applies under early stopping.

Theorem A.15 (τ-leaping sampler guarantee)**.** *For a fixed grid $\mathcal{T} = \{\delta = t_0 < t_1 < \cdots < t_N = T\}$, let $\Delta := \max_{1\le k\le N}(t_k - t_{k-1})$ and suppose $\Delta \lesssim 1$. Then the τ-leaping sampler initialized from π satisfies*

$$D_{\mathrm{KL}}\left(P_\delta \| Q_\delta^{\tau,\mathcal{T},\sigma}\right) \lesssim \mathcal{L}_{\mathcal{T}}^{\mathrm{SE}}(\sigma;P_0) + e^{-T} d\log K + \Delta d\log\left(\frac{K}{\Delta}\right).$$

A.5 ReLU Network Constructions

We first record two ReLU network constructions following Oko et al. (2023, Lemmas F.6–F.7). For multiplication, we use the bounded-input construction in the proof of Lemma F.6, restricted to $[0,1]^d$ and clipped to $[0,1]$. We omit the proofs.

Lemma A.16 (ReLU multiplication network)**.** *For every integer $d \ge 2$ and $\epsilon \in (0,1)$, there exists a ReLU network $\mathrm{Mult}_{d,\epsilon} : [0,1]^d \to [0,1]$ such that*

$$\sup_{u\in[0,1]^d} \left|\mathrm{Mult}_{d,\epsilon}(u) - \prod_{j=1}^d u_j\right| \le \epsilon.$$

Moreover, the network can be chosen with

$$L_{\mathrm{mult}} \lesssim \log d\big(\log\epsilon^{-1} + \log d\big), \qquad W_{\mathrm{mult}} \lesssim d, \qquad S_{\mathrm{mult}} \lesssim d\big(\log\epsilon^{-1} + \log d\big), \qquad B_{\mathrm{mult}} \lesssim 1.$$

Lemma A.17 (ReLU reciprocal network)**.** *For every $\epsilon \in (0,1)$, there exists a ReLU network $\mathrm{Rec}_\epsilon : \mathbb{R} \to \mathbb{R}$ such that, for every $u \in [\epsilon, \epsilon^{-1}]$ and $u' \in \mathbb{R}$,*

$$\left|\mathrm{Rec}_\epsilon(u') - \frac{1}{u}\right| \le \epsilon + \frac{|u'-u|}{\epsilon^2}.$$

Moreover, the network can be chosen with

$$L_{\mathrm{rec}} \lesssim \log^2\epsilon^{-1}, \qquad W_{\mathrm{rec}} \lesssim \log^3\epsilon^{-1}, \qquad S_{\mathrm{rec}} \lesssim \log^4\epsilon^{-1}, \qquad B_{\mathrm{rec}} \lesssim \epsilon^{-2}.$$

Using Lemmas A.16 and A.17, we construct a ReLU network that approximates the discrete score simultaneously over a fixed time grid.

Lemma A.18 (ReLU approximation of discrete scores on a fixed grid)**.** *Let $\mathcal{T} = \{\delta = t_0 < t_1 < \cdots < t_N = T\}$ be a fixed grid, and let $\mu = \sum_{v=1}^D w_v \delta_{\boldsymbol{z}_v}$ with $w_v > 0$ and $\sum_{v=1}^D w_v = 1$. Write $\mu_t := \mu\boldsymbol{P}_t$. For every $\varepsilon \in (0,1)$, there exists a ReLU network $G_\varepsilon : \mathbb{R}^{1+2dK} \to \mathbb{R}$ with range contained in $[B_1(t_1)^{-1}, B_1(t_1)]$ such that*

$$\max_{1\le k\le N} \max_{d_{\mathrm{Ham}}(\boldsymbol{x},\boldsymbol{y})=1} \left|G_\varepsilon\left(\frac{k}{N}, \boldsymbol{e}(\boldsymbol{y}), \boldsymbol{e}(\boldsymbol{x})\right) - \frac{\mu_{t_k}(\boldsymbol{y})}{\mu_{t_k}(\boldsymbol{x})}\right| \le \varepsilon.$$

The network may be chosen with

$$L \lesssim \log^2\left(\frac{B_0(t_1)}{\varepsilon}\right), \qquad W \lesssim d(KN+D) + \log^3\left(\frac{B_0(t_1)}{\varepsilon}\right),$$

$$S \lesssim dK^2N + dD\log\left(\frac{B_0(t_1)}{\varepsilon}\right) + \log^4\left(\frac{B_0(t_1)}{\varepsilon}\right), \qquad B \lesssim N + B_0(t_1)^2 + \varepsilon^{-2}.$$

Proof. We first approximate the two forward marginals, then approximate the reciprocal of the denominator, and finally multiply.

Step 1: Time and state selection. For $k \in [N]$, define

$$J_k(r) := \rho\left(1 - N\rho\left(r - \frac{k}{N}\right) - N\rho\left(\frac{k}{N} - r\right)\right),$$

so that $J_k(\ell/N) = \mathbf{1}\{k = \ell\}$. For $i \in [d]$ and $a \in [K]$, define $I^{\boldsymbol{x}}_{k,i,a}(r, \boldsymbol{e}(\boldsymbol{x})) := \rho(J_k(r) + (\boldsymbol{e}(\boldsymbol{x}))_{i,a} - 1)$, and define $I^{\boldsymbol{y}}_{k,i,a}$ analogously. At encoded inputs, $I^{\boldsymbol{x}}_{k,i,a}(\ell/N, \boldsymbol{e}(\boldsymbol{x})) = \mathbf{1}\{k = \ell, x^i = a\}$.

Step 2: Coordinatewise transition probabilities. For $i \in [d]$ and $b \in [K]$, define

$$G^{\boldsymbol{x}}_{i,b} := \sum_{k=1}^{N}\sum_{a=1}^{K} \boldsymbol{P}^{\mathrm{coord}}_{t_k}(b,a) I^{\boldsymbol{x}}_{k,i,a},$$

so that $G^{\boldsymbol{x}}_{i,b}(k/N, \boldsymbol{e}(\boldsymbol{x})) = \boldsymbol{P}^{\mathrm{coord}}_{t_k}(b, x^i)$, and define $G^{\boldsymbol{y}}_{i,b}$ analogously.

Step 3: Forward marginals. Set $\eta := \min\{B_0(t_1)^{-1}, \varepsilon/16\}$ and $\varepsilon_1 := \varepsilon\eta^2/16$. For each $v \in [D]$, Lemma A.16 applied to $G^{\boldsymbol{x}}_{1,z^1_v}, \ldots, G^{\boldsymbol{x}}_{d,z^d_v}$ gives an ε_1-approximation to $\boldsymbol{P}_{t_k}(\boldsymbol{z}_v, \boldsymbol{x})$ at every encoded grid input. Taking the weighted sum over v produces networks $F_{\boldsymbol{x}}$ and $F_{\boldsymbol{y}}$ satisfying

$$|F_{\boldsymbol{x}} - \mu_{t_k}(\boldsymbol{x})| \le \varepsilon_1, \qquad |F_{\boldsymbol{y}} - \mu_{t_k}(\boldsymbol{y})| \le \varepsilon_1$$

at every encoded input.

Step 4: Reciprocal. Clip $F_{\boldsymbol{x}}$ to $[B_0(t_1)^{-1}, 1]$, which cannot increase its error by Lemma A.4. Since $\eta \le \mu_{t_k}(\boldsymbol{x}) \le 1 \le \eta^{-1}$, Lemma A.17, followed by clipping to $[1, B_0(t_1)]$, gives a network $R_{\boldsymbol{x}}$ such that

$$\left|R_{\boldsymbol{x}} - \mu_{t_k}(\boldsymbol{x})^{-1}\right| \le \eta + \frac{\varepsilon_1}{\eta^2} \le \frac{\varepsilon}{8}.$$

Step 5: Score ratio. Clip $F_{\boldsymbol{y}}$ to $[0,1]$, apply $\mathrm{Mult}_{2,\varepsilon/(4B_0(t_1))}$ to $F_{\boldsymbol{y}}$ and $R_{\boldsymbol{x}}/B_0(t_1)$, and multiply the output by $B_0(t_1)$. The resulting network $\widetilde{G}_\varepsilon$ satisfies

$$\left|\widetilde{G}_\varepsilon - \frac{\mu_{t_k}(\boldsymbol{y})}{\mu_{t_k}(\boldsymbol{x})}\right| \le \frac{\varepsilon}{4} + B_0(t_1)\varepsilon_1 + \frac{\varepsilon}{8} < \varepsilon,$$

because $\eta \le B_0(t_1)^{-1}$ implies $B_0(t_1)\varepsilon_1 \le \varepsilon/16$. Finally, define $G_\varepsilon := \mathrm{clip}_{[B_1(t_1)^{-1}, B_1(t_1)]}(\widetilde{G}_\varepsilon)$. Since B_1 is decreasing and $t_k \ge t_1$, Lemma A.4 shows that the true score lies in this interval, so the final clipping cannot increase the error.

It remains to collect the architecture bounds. The selector and transition layers have constant depth, width $O(dKN)$, sparsity $O(dK^2N)$, and weights of order N. The $2D$ multiplication networks contribute width $O(dD)$ and sparsity $O(dD(\log \varepsilon_1^{-1} + \log d))$, while the reciprocal contributes depth $O(\log^2 \eta^{-1})$, width $O(\log^3 \eta^{-1})$, and sparsity $O(\log^4 \eta^{-1})$. Since $B_0(t_1) \ge K^d$, $\eta^{-1} \le 16B_0(t_1)/\varepsilon$, and $\log \varepsilon_1^{-1} \lesssim \log(B_0(t_1)/\varepsilon)$, these contributions give the stated bounds on L, W, S. Finally, the largest weights arise from the time selectors and reciprocal network, and $\eta^{-2} \lesssim B_0(t_1)^2 + \varepsilon^{-2}$, yielding $B \lesssim N + B_0(t_1)^2 + \varepsilon^{-2}$. □

B Proofs for TV Results

B.1 Upper Bound: Proofs of Theorem 3 and Corollary 4

Proof of Theorem 3. Fix a realization of the observations, so that $\widehat{P}_0$ is supported on D_n points. Recall that $M = B_1(t_1)$, and set

$$\varepsilon_{\mathrm{TV}} := \sqrt{\frac{\mathfrak{s}_n(P_0)}{dnMT}}.$$

Since $\mathfrak{s}_n(P_0) \le n$, $d \ge 2$, $M \ge 1$, and $T \ge 1$, we have $\varepsilon_{\mathrm{TV}} \in (0,1)$. Applying Lemma A.18 to $\widehat{P}_0$ gives a score $\sigma^\dagger$ such that

$$\max_{1\le k\le N} \max_{d_{\mathrm{Ham}}(\boldsymbol{x},\boldsymbol{y})=1} \left|\sigma^\dagger_{t_k}(\boldsymbol{y},\boldsymbol{x}) - \widehat{\sigma}^\star_{t_k}(\boldsymbol{y},\boldsymbol{x})\right| \le \varepsilon_{\mathrm{TV}},$$

where $\widehat{\sigma}^\star_{t_k}(\boldsymbol{y},\boldsymbol{x}) := \widehat{P}_{t_k}(\boldsymbol{y})/\widehat{P}_{t_k}(\boldsymbol{x})$.

Both $\sigma^\dagger$ and $\widehat{\sigma}^\star$ take values in $[M^{-1}, M]$. Since $f''(u) = 1/u$, Taylor's theorem gives $D_{\mathrm{Br}}(u,v) \le M(u-v)^2/2$ for $u, v \in [M^{-1}, M]$. Moreover, $\sum_{\boldsymbol{y}:\boldsymbol{Q}(\boldsymbol{y},\boldsymbol{x})>0} \boldsymbol{Q}(\boldsymbol{y},\boldsymbol{x}) = d(K-1)/K \le d$, and hence

$$\mathcal{L}^{\mathrm{SE}}_{\mathcal{T}}(\sigma^\dagger; \widehat{P}_0) \le \frac{dM\varepsilon^2_{\mathrm{TV}}}{2} \sum_{k=1}^{N} (t_k - t_{k-1}) \le \frac{dMT\varepsilon^2_{\mathrm{TV}}}{2} = \frac{\mathfrak{s}_n(P_0)}{2n}.$$

By Lemma A.14 and the $\varepsilon_{\mathrm{opt}}$-optimality of $\widehat{\sigma}$,

$$\mathcal{L}^{\mathrm{SE}}_{\mathcal{T}}(\widehat{\sigma}; \widehat{P}_0) \le \mathcal{L}^{\mathrm{SE}}_{\mathcal{T}}(\sigma^\dagger; \widehat{P}_0) + \varepsilon_{\mathrm{opt}} \lesssim \frac{\mathfrak{s}_n(P_0)}{n}.$$

Applying Assumption 2 conditionally on the observations and using the assumed sampler remainder gives

$$D_{\mathrm{KL}}\left(\widehat{P}_\delta \| Q^{\mathcal{A}}_\delta\right) \lesssim \frac{\mathfrak{s}_n(P_0)}{n}.$$

Thus, by Pinsker's inequality, $d_{\mathrm{TV}}(\widehat{P}_\delta, Q^{\mathcal{A}}_\delta) \lesssim \sqrt{\mathfrak{s}_n(P_0)/n}$.

By coupling the forward process initialized from $\widehat{P}_0$ with its initial state,

$$d_{\mathrm{TV}}(\widehat{P}_0, \widehat{P}_\delta) \le 1 - \exp\left(-\frac{d(K-1)}{K}\delta\right) \le d\delta.$$

Therefore, the triangle inequality, Lemma A.6, and the assumption on δ yield

$$\mathbb{E}_{P_0} d_{\mathrm{TV}}\left(P_0, Q^{\mathcal{A}}_\delta\right) \lesssim \sqrt{\frac{\mathfrak{s}_n(P_0)}{n}} + d\delta \lesssim \sqrt{\frac{\mathfrak{s}_n(P_0)}{n}}.$$

It remains to verify the architecture bounds. By Lemma A.18, it suffices to control $B_0(t_1)/\varepsilon_{\mathrm{TV}}$. Since $B_0(t_1) = (K/(1-e^{-t_1}))^d$ and $M = B_1(t_1) \le K/(1-e^{-t_1})$,

$$\log\left(\frac{B_0(t_1)}{\varepsilon_{\mathrm{TV}}}\right) \lesssim d \log\left(\frac{K}{1-e^{-t_1}}\right) + \log\left(\frac{nT}{\mathfrak{s}_n(P_0)}\right),$$

where we used $K/(1-e^{-t_1}) \ge 2$, $\mathfrak{s}_n(P_0) \le n$, and $T \ge 1$. Substituting this bound into Lemma A.18 gives the stated bounds on L, W, S. Finally,

$$B_0(t_1)^2 = \left(\frac{K}{1-e^{-t_1}}\right)^{2d}, \qquad \varepsilon^{-2}_{\mathrm{TV}} = \frac{dnMT}{\mathfrak{s}_n(P_0)},$$

which gives the stated bound on B. □

Proof of Corollary 4. Let $\Lambda := \log(Kdn/\mathfrak{s}_n(P_0))$. Since $\mathfrak{s}_n(P_0) \le n$ and $K, d \ge 2$, we have $\Lambda > 1$.

For the stated grid, $\Delta = \mathfrak{s}_n(P_0)/(dn\Lambda)$. By Theorem A.15,

$$\mathcal{R}_\tau(\mathcal{T}, \theta; M) = e^{-T} d \log K + \Delta d \log\left(\frac{K}{\Delta}\right).$$

The definitions of T and Δ give

$$e^{-T} d \log K = \frac{\mathfrak{s}_n(P_0)}{en}, \qquad \Delta d \log\left(\frac{K}{\Delta}\right) = \frac{\mathfrak{s}_n(P_0)}{n}\left(1 + \frac{\log \Lambda}{\Lambda}\right) \lesssim \frac{\mathfrak{s}_n(P_0)}{n}.$$

Hence $\mathcal{R}_\tau(\mathcal{T}, \theta; M) \lesssim \mathfrak{s}_n(P_0)/n$. Moreover, $\delta = 0$ and $T \geq \log(2e \log 2) > 1$, so Theorem 3 applies.

It remains to simplify the architecture bounds. Since $t_1 = \Delta \leq 1/(d\Lambda) < 1$, the inequality $1 - e^{-\Delta} \geq \Delta e^{-\Delta}$ gives

$$\frac{K}{1 - e^{-\Delta}} \leq \frac{eKdn\Lambda}{\mathfrak{s}_n(P_0)}, \qquad d\Delta \leq \frac{1}{\Lambda}.$$

Therefore,

$$\log\left(\frac{K}{1 - e^{-\Delta}}\right) \lesssim \Lambda.$$

Also, since $\log K \leq K/e$,

$$T = \Lambda + \log\left(\frac{e \log K}{K}\right) \leq \Lambda.$$

Consequently,

$$N = \left\lceil \frac{T}{\Delta} \right\rceil \lesssim \frac{dn}{\mathfrak{s}_n(P_0)}\Lambda^2, \qquad \log\left(\frac{nT}{\mathfrak{s}_n(P_0)}\right) \lesssim \Lambda.$$

Substituting these bounds into Theorem 3 yields

$$L \lesssim d^2\Lambda^2,$$

$$W \lesssim \frac{d^2Kn}{\mathfrak{s}_n(P_0)}\Lambda^2 + dD_n + d^3\Lambda^3,$$

and

$$S \lesssim \frac{d^2K^2n}{\mathfrak{s}_n(P_0)}\Lambda^2 + d^2D_n\Lambda + d^4\Lambda^4.$$

Finally, $M = B_1(\Delta) \leq K/(1 - e^{-\Delta}) \lesssim Kdn\Lambda/\mathfrak{s}_n(P_0)$, and

$$\left(\frac{K}{1 - e^{-\Delta}}\right)^{2d} \lesssim \left(\frac{Kdn\Lambda}{\mathfrak{s}_n(P_0)}\right)^{2d},$$

where we used $d\Delta \leq 1/\Lambda$. Since $T \leq \Lambda$, the remaining terms N and $dnMT/\mathfrak{s}_n(P_0)$ are dominated by the same quantity. Thus

$$B \lesssim \left[\frac{Kdn}{\mathfrak{s}_n(P_0)} \log\left(\frac{Kdn}{\mathfrak{s}_n(P_0)}\right)\right]^{2d}.$$

□

B.2 Minimax Lower Bound: Proof of Theorem 5

Proof of Theorem 5. The upper bound follows immediately from Lemma A.6 and $\mathfrak{s}_n(P_0) \leq s_n$ for $P_0 \in \mathcal{P}_{s_n}$.

For the lower bound, let

$$m := \left\lfloor \frac{s_n}{2} \right\rfloor, \qquad \gamma := \frac{1}{2\sqrt{2}}\sqrt{\frac{m}{n}}.$$

Since $2 \leq s_n \leq n$, we have $0 < \gamma \leq 1/4$, and since $2m \leq s_n \leq K^d$, choose distinct states $\boldsymbol{x}_1, \boldsymbol{y}_1, \ldots, \boldsymbol{x}_m, \boldsymbol{y}_m \in [K]^d$. For each $\theta \in \{-1, 1\}^m$, define $P_0^\theta \in \Delta_{K^d}$ by

$$P_0^\theta(\boldsymbol{x}_i) := \frac{1 + \gamma\theta_i}{2m}, \qquad P_0^\theta(\boldsymbol{y}_i) := \frac{1 - \gamma\theta_i}{2m}, \qquad i \in [m],$$

and set $P_0^\theta(\boldsymbol{z}) = 0$ elsewhere. Since $\mathfrak{s}_n(P_0^\theta) \le |\operatorname{supp}(P_0^\theta)| = 2m \le s_n$, every P_0^θ belongs to $\mathcal{P}_{s_n}$. Fix any possibly randomized estimator $\widetilde{P}$, and define $\widetilde{\theta}_i = 1$ if $\widetilde{P}(\boldsymbol{x}_i) \ge \widetilde{P}(\boldsymbol{y}_i)$ and $\widetilde{\theta}_i = -1$ otherwise. By the triangle inequality,

$$d_{\mathrm{TV}}(\widetilde{P}, P_0^\theta) \ge \frac{\gamma}{2m} \sum_{i=1}^{m} \mathbf{1}\{\widetilde{\theta}_i \ne \theta_i\}.$$

Let $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$ denote probability and expectation under $(P_0^\theta)^{\otimes n}$, including any auxiliary randomness of $\widetilde{P}$. Averaging uniformly over θ gives

$$\sup_\theta \mathbb{E}_\theta d_{\mathrm{TV}}(\widetilde{P}, P_0^\theta) \ge \frac{\gamma}{2m} \sum_{i=1}^{m} \frac{1}{2^m} \sum_{\theta} \mathbb{P}_\theta(\widetilde{\theta}_i \ne \theta_i).$$

Fix $i \in [m]$ and $\theta_{-i} \in \{-1,1\}^{m-1}$, and let P_0^+ and P_0^- denote the distributions obtained by setting $\theta_i = 1$ and $\theta_i = -1$, respectively. A direct calculation gives

$$D_{\mathrm{KL}}(P_0^+ \| P_0^-) = \frac{\gamma}{m} \log\left(\frac{1+\gamma}{1-\gamma}\right) \le \frac{4\gamma^2}{m},$$

where we used $\log((1+t)/(1-t)) \le 4t$ for $0 \le t \le 1/2$. Therefore,

$$D_{\mathrm{KL}}\left((P_0^+)^{\otimes n} \| (P_0^-)^{\otimes n}\right) \le \frac{1}{2},$$

and Pinsker's inequality gives $d_{\mathrm{TV}}((P_0^+)^{\otimes n}, (P_0^-)^{\otimes n}) \le 1/2$. Le Cam's two-point inequality then yields

$$\frac{1}{2}\left\{\mathbb{P}_+(\widetilde{\theta}_i = -1) + \mathbb{P}_-(\widetilde{\theta}_i = 1)\right\} \ge \frac{1}{4}.$$

Averaging over θ_{-i} shows that

$$\frac{1}{2^m} \sum_{\theta} \mathbb{P}_\theta(\widetilde{\theta}_i \ne \theta_i) \ge \frac{1}{4}.$$

Consequently,

$$\sup_\theta \mathbb{E}_\theta d_{\mathrm{TV}}(\widetilde{P}, P_0^\theta) \ge \frac{\gamma}{8} \ge \frac{1}{32\sqrt{2}} \sqrt{\frac{s_n}{n}},$$

where we used $m = \lfloor s_n/2 \rfloor \ge s_n/4$.

Since $\{P_0^\theta : \theta \in \{-1,1\}^m\} \subseteq \mathcal{P}_{s_n}$, every estimator $\widetilde{P}$ satisfies

$$\sup_{P_0 \in \mathcal{P}_{s_n}} \mathbb{E}_{P_0} d_{\mathrm{TV}}(P_0, \widetilde{P}) \ge \frac{1}{32\sqrt{2}} \sqrt{\frac{s_n}{n}}.$$

Taking the infimum over all estimators $\widetilde{P}$ completes the proof. □

C Proofs for KL Results

C.1 Upper Bound: Proofs of Theorem 6 and Corollary 7

Proof of Theorem 6. The proof follows that of Theorem 3, except that Lemma A.13 replaces the triangle inequality. Set

$$\varepsilon_{\mathrm{KL}} := \frac{1}{\sqrt{dMTB_0(\delta)}}.$$

Since $d, M, T, B_0(\delta) \ge 1$, we have $\varepsilon_{\mathrm{KL}} \in (0,1)$. Applying Lemma A.18 to $\widehat{P}_0$ and arguing as in the proof of Theorem 3 gives a clipped score $\sigma^\dagger$ satisfying

$$\mathcal{L}_{\mathcal{T}}^{\mathrm{SE}}(\sigma^\dagger; \widehat{P}_0) \le \frac{dMT\varepsilon_{\mathrm{KL}}^2}{2} = \frac{1}{2B_0(\delta)}.$$

Hence, by Lemma A.14 and the $\varepsilon_{\mathrm{opt}}$-optimality of $\widehat{\sigma}$,

$$\mathcal{L}^{\mathrm{SE}}_{\mathcal{T}}(\widehat{\sigma};\widehat{P}_0) \lesssim \frac{1}{B_0(\delta)}.$$

Applying Assumption 2 conditionally on the observations and using the assumed sampler remainder gives

$$D_{\mathrm{KL}}\left(\widehat{P}_\delta \| Q^{\mathcal{A}}_\delta\right) \lesssim \frac{1}{B_0(\delta)}.$$

By Lemma A.4, $\widehat{P}_\delta(\boldsymbol{x}) \geq B_0(\delta)^{-1}$ for every $\boldsymbol{x}$. Therefore, Lemma A.13 yields

$$D_{\mathrm{KL}}\left(P_0 \| Q^{\mathcal{A}}_\delta\right) \lesssim D_{\mathrm{KL}}(P_0 \| \widehat{P}_\delta) + \frac{1}{B_0(\delta)}.$$

Since $\delta = 1/n$, Lemma A.11 controls the first term. Moreover,

$$B_0(\delta)^{-1} = \left(\frac{1-e^{-1/n}}{K}\right)^d \leq (Kn)^{-d} \lesssim \frac{\log n}{n} \leq \frac{\mathfrak{s}_n(P_0)\log(eK^d/\mathfrak{s}_n(P_0))\log n}{n}.$$

Taking expectations therefore gives

$$\mathbb{E}_{P_0} D_{\mathrm{KL}}\left(P_0 \| Q^{\mathcal{A}}_\delta\right) \lesssim \frac{\mathfrak{s}_n(P_0)\log\left(eK^d/\mathfrak{s}_n(P_0)\right)\log n}{n}.$$

It remains to verify the architecture bounds. Since $t_1 > \delta = 1/n$ and $1 - e^{-1/n} \geq 1/(2n)$,

$$B_0(t_1) \leq B_0(\delta) \leq (2Kn)^d, \qquad M \leq \frac{K}{1-e^{-1/n}} \leq 2Kn.$$

Thus

$$\log\left(\frac{B_0(t_1)}{\varepsilon_{\mathrm{KL}}}\right) \lesssim d\log(Kn) + \log T.$$

Substituting this into Lemma A.18 gives the stated bounds on L, W, S. Finally,

$$B_0(t_1)^2 = \left(\frac{K}{1-e^{-t_1}}\right)^{2d}, \qquad \varepsilon_{\mathrm{KL}}^{-2} = dMT\left(\frac{K}{1-e^{-1/n}}\right)^d,$$

which gives the stated bound on B. □

Proof of Corollary 7. Define

$$\Xi := \frac{K}{1-e^{-1/n}}.$$

Since $1 - e^{-1/n} \leq 1/n$, we have $\Xi \geq Kn$. First note that $T = \log(ed\Xi^d \log K) \geq 1$.

For the stated grid, $\Delta = (d^2\Xi^d \log(Kn))^{-1}$. By Theorem A.15,

$$\mathcal{R}_\tau(\mathcal{T},\theta;M) \lesssim e^{-T} d\log K + \Delta d \log\left(\frac{K}{\Delta}\right).$$

Since

$$e^{-T} d\log K = \frac{1}{e\Xi^d}, \qquad \log\left(\frac{K}{\Delta}\right) = \log\left(d^2 K \Xi^d \log(Kn)\right) \lesssim d\log(Kn),$$

we obtain $\mathcal{R}_\tau(\mathcal{T},\theta;M) \lesssim \Xi^{-d}$, as required by Theorem 6.

It remains to simplify the architecture bounds. Since $t_1 = \delta + \Delta > \delta = 1/n$,

$$\frac{K}{1-e^{-t_1}} \leq \Xi, \qquad M = B_1(t_1) \leq \Xi.$$

Moreover, $T \lesssim d\log(Kn)$, and hence

$$N = \left\lceil \frac{T-\delta}{\Delta} \right\rceil \lesssim d^3 \Xi^d \log^2(Kn).$$

Substituting these bounds into Theorem 6, and using $D_n \leq n \leq \Xi^d$, gives

$$L \lesssim d^2 \log^2(Kn), \qquad W \lesssim d^4 K \Xi^d \log^2(Kn), \qquad S \lesssim d^4 K^2 \Xi^d \log^2(Kn).$$

Finally, the bound on B in Theorem 6, together with $M \leq \Xi$, $T \lesssim d\log(Kn)$, and the bound on N, gives $B \lesssim \Xi^{2d}$. □

C.2 Minimax Lower Bound: Proof of Theorem 8

The empirical distribution $\widehat{P}_0$ assigns zero probability to unobserved states and therefore has infinite KL risk. We instead use the adaptive add-λ estimator analyzed by Mourtada (2026), specialized to the state space $[K]^d$.

Lemma C.1. *Define the adaptive additive-smoothing estimator*

$$\widehat{P}^{\mathrm{ad}}(\boldsymbol{x}) := \frac{N_{\boldsymbol{x}} + D_n/K^d}{n + D_n}, \qquad \boldsymbol{x} \in [K]^d.$$

Then

$$\mathbb{E}_{P_0} D_{\mathrm{KL}}\left(P_0 \| \widehat{P}^{\mathrm{ad}}\right) \le 4.4\, \frac{\mathfrak{s}_n(P_0) \log\left(eK^d/\mathfrak{s}_n(P_0)\right)}{n}.$$

Proof. For $r > 0$, define

$$\mathfrak{s}^{\circ}_r(P_0) := \sum_{\boldsymbol{x} \in [K]^d} \min\left\{e^{1-rP_0(\boldsymbol{x})}, rP_0(\boldsymbol{x})\right\}.$$

By Mourtada (2026, Proposition 2), specialized to an alphabet of size K^d,

$$\mathbb{E}_{P_0} D_{\mathrm{KL}}\left(P_0 \| \widehat{P}^{\mathrm{ad}}\right) \le \frac{2.4\mathfrak{s}_n(P_0) + 2\mathfrak{s}^{\circ}_{n/2}(P_0) \log\left(eK^d/\mathfrak{s}^{\circ}_{n/2}(P_0)\right)}{n+1}.$$

For every $u \ge 0$,

$$\min\left\{e^{1-u/2}, \frac{u}{2}\right\} \le \min\{u, 1\},$$

as follows by considering $u \le 1$, $1 \le u \le 2$, and $u \ge 2$. Applying this pointwise with $u = nP_0(\boldsymbol{x})$ gives $\mathfrak{s}^{\circ}_{n/2}(P_0) \le \mathfrak{s}_n(P_0)$. Since $t \mapsto t\log(eK^d/t)$ is increasing on $(0, K^d]$ and $\mathfrak{s}_n(P_0) \le K^d$,

$$\mathfrak{s}^{\circ}_{n/2}(P_0) \log\left(\frac{eK^d}{\mathfrak{s}^{\circ}_{n/2}(P_0)}\right) \le \mathfrak{s}_n(P_0) \log\left(\frac{eK^d}{\mathfrak{s}_n(P_0)}\right), \qquad \mathfrak{s}_n(P_0) \le \mathfrak{s}_n(P_0) \log\left(\frac{eK^d}{\mathfrak{s}_n(P_0)}\right).$$

Substituting these bounds and using $n + 1 \ge n$ proves the claim. □

Now we prove that rate in Lemma C.1 is actually minimax risk.

Proof of Theorem 8. Since $t \mapsto t\log(eK^d/t)$ is increasing on $(0, K^d]$, the upper bound follows from Lemma C.1.

For the lower bound, set

$$m := \min\left\{\lfloor s_n \rfloor, \left\lfloor \frac{K^d}{2} \right\rfloor\right\}.$$

Since $2 \le s_n \le \min\{n, K^d\}$, we have $2 \le m \le n$ and $2m \le K^d$. Fix $\boldsymbol{x}_\star \in [K]^d$, and let

$$\mathcal{J} := \left\{J \subseteq [K]^d \setminus \{\boldsymbol{x}_\star\} : |J| = m - 1\right\}.$$

For each $J \in \mathcal{J}$, define P_0^J by

$$P_0^J(\boldsymbol{x}_\star) := 1 - \frac{m-1}{n}, \qquad P_0^J(\boldsymbol{x}) := \frac{1}{n}\mathbf{1}\{\boldsymbol{x} \in J\}, \quad \boldsymbol{x} \ne \boldsymbol{x}_\star.$$

Since $\mathfrak{s}_n(P_0^J) \le |\operatorname{supp}(P_0^J)| = m \le s_n$, every P_0^J belongs to $\mathcal{P}_{s_n}$.

Draw J uniformly from $\mathcal{J}$. For any possibly randomized estimator $\widetilde{P}$,

$$\sup_{P_0 \in \mathcal{P}_{s_n}} \mathbb{E}_{P_0} D_{\mathrm{KL}}(P_0 \| \widetilde{P}) \ge \mathbb{E}_J \mathbb{E}_{P_0^J} D_{\mathrm{KL}}(P_0^J \| \widetilde{P}).$$

Let $O := \{\boldsymbol{x} \ne \boldsymbol{x}_\star : N_{\boldsymbol{x}} \ge 1\}$ be the set of observed rare states. Under P_0^J, we have $O \subseteq J$, and the likelihood under $P_0^{J'}$ is proportional to $\mathbf{1}\{O \subseteq J'\}$. Hence, conditional on the sample, the posterior distribution of J is uniform over the sets $J' \in \mathcal{J}$ containing O.

Let $\overline{P} := \mathbb{E}[P_0^J \mid X_0^{(1)}, \ldots, X_0^{(n)}]$. Then

$$\overline{P}(\boldsymbol{x}) = \begin{cases} 1 - \frac{m-1}{n}, & \boldsymbol{x} = \boldsymbol{x}_\star, \\ \frac{1}{n}, & \boldsymbol{x} \in O, \\ \frac{1}{n}\frac{m-1-|O|}{K^d-1-|O|}, & \boldsymbol{x} \notin O \cup \{\boldsymbol{x}_\star\}. \end{cases}$$

For every $Q \in \Delta_{K^d}$,

$$\mathbb{E}\Big[D_{\mathrm{KL}}(P_0^J \| Q) \mid X_0^{(1)}, \ldots, X_0^{(n)}\Big] = C(X_0^{(1)}, \ldots, X_0^{(n)}) + D_{\mathrm{KL}}(\overline{P} \| Q),$$

where C does not depend on Q. Thus the conditional Bayes risk is minimized by $Q = \overline{P}$, and the same conclusion holds for randomized estimators after averaging over their auxiliary randomness. Consequently,

$$\mathbb{E}_J \mathbb{E}_{P_0^J} D_{\mathrm{KL}}(P_0^J \| \widetilde{P}) \geq \mathbb{E}_J \mathbb{E}_{P_0^J} D_{\mathrm{KL}}(P_0^J \| \overline{P}).$$

If $m - 1 - |O| > 0$, only the unobserved states in $J \setminus O$ contribute, and

$$D_{\mathrm{KL}}(P_0^J \| \overline{P}) = \frac{m-1-|O|}{n} \log\left(\frac{K^d - 1 - |O|}{m-1-|O|}\right) \geq \frac{m-1-|O|}{n} \log\left(\frac{K^d}{m}\right).$$

The same lower bound holds trivially when $m - 1 - |O| = 0$. Moreover,

$$\mathbb{E}_{P_0^J}[m - 1 - |O|] = (m-1)\left(1 - \frac{1}{n}\right)^n.$$

Therefore,

$$\mathbb{E}_J \mathbb{E}_{P_0^J} D_{\mathrm{KL}}(P_0^J \| \overline{P}) \geq \frac{m-1}{n}\left(1 - \frac{1}{n}\right)^n \log\left(\frac{K^d}{m}\right).$$

Since $n \geq 2$, $m \geq 2$, and $K^d \geq 2m$,

$$\left(1 - \frac{1}{n}\right)^n \geq \frac{1}{4}, \qquad m - 1 \geq \frac{m}{2}, \qquad \log\left(\frac{K^d}{m}\right) \geq \frac{1}{3}\log\left(\frac{eK^d}{m}\right).$$

Hence

$$\mathbb{E}_J \mathbb{E}_{P_0^J} D_{\mathrm{KL}}(P_0^J \| \widetilde{P}) \geq \frac{1}{24} \frac{m \log(eK^d/m)}{n}.$$

It remains to compare m with s_n. If $s_n \leq K^d/2$, then $m = \lfloor s_n \rfloor \geq s_n/2$, so

$$m \log\left(\frac{eK^d}{m}\right) \geq \frac{s_n}{2} \log\left(\frac{eK^d}{s_n}\right).$$

If $s_n > K^d/2$, then $m = \lfloor K^d/2 \rfloor \geq K^d/3$, while $s_n \log(eK^d/s_n) \leq K^d$, and hence

$$m \log\left(\frac{eK^d}{m}\right) \geq \frac{1}{3} s_n \log\left(\frac{eK^d}{s_n}\right).$$

Thus, in both cases,

$$m \log\left(\frac{eK^d}{m}\right) \geq \frac{1}{3} s_n \log\left(\frac{eK^d}{s_n}\right).$$

Since $\widetilde{P}$ was arbitrary,

$$\inf_{\widetilde{P}} \sup_{P_0 \in \mathcal{P}_{s_n}} \mathbb{E}_{P_0} D_{\mathrm{KL}}(P_0 \| \widetilde{P}) \geq \frac{1}{72} \frac{s_n \log(eK^d/s_n)}{n}.$$

□

D EXACT ONE-STEP EMPIRICAL REVERSE SAMPLING

We construct an exact one-step empirical reverse sampler that attains the same TV and KL rates as the fixed-grid samplers in Theorems 3 and 6, while requiring only $O(d(D_n + K))$ operations per generated sample once the distinct observations and their multiplicities have been stored. This shows that the unfavorable computational dependence in Corollary 7 is not intrinsic to the statistical KL rate. Rather than discretizing reverse time or simulating intermediate jumps, the sampler draws directly from a finite-time empirical reverse kernel.

D.1 EXACT REVERSE KERNEL AND SAMPLING ALGORITHM

For $0 \le \delta < T$, define the empirical reverse transition kernel from forward time T to δ by

$$\overleftarrow{\boldsymbol{P}}_{T\to\delta,\widehat{P}}(\boldsymbol{x},\boldsymbol{y}) := \frac{\widehat{P}_\delta(\boldsymbol{y})\boldsymbol{P}_{T-\delta}(\boldsymbol{y},\boldsymbol{x})}{\widehat{P}_T(\boldsymbol{x})}, \qquad \boldsymbol{x},\boldsymbol{y}\in[K]^d. \tag{D.1}$$

By Lemma A.14, the unrestricted minimizer of the exact empirical denoising score-entropy loss is the empirical score $\widehat{\sigma}_t^\star(\boldsymbol{y},\boldsymbol{x}) = \widehat{P}_t(\boldsymbol{y})/\widehat{P}_t(\boldsymbol{x})$. If this score is available for every $t\in[\delta,T]$ and the reverse dynamics are simulated exactly, the resulting reverse rates are those of the time reversal of the forward process initialized from $\widehat{P}_0$. Bayes' rule then gives (D.1) as its finite-time transition kernel. We initialize this kernel from the stationary distribution π, and denote the resulting output distribution by

$$Q_\delta^{\mathrm{os}} := \pi\overleftarrow{\boldsymbol{P}}_{T\to\delta,\widehat{P}}.$$

Thus, relative to the exact empirical reverse process, the only initialization discrepancy is replacing $\widehat{P}_T$ by π.

Write $\mathrm{supp}(\widehat{P}_0) = \{\boldsymbol{z}_1,\dots,\boldsymbol{z}_{D_n}\}$, so that $\widehat{P}_0(\boldsymbol{z}_i) = N_{\boldsymbol{z}_i}/n$. Given $X_T = \boldsymbol{x}$, Bayes' rule gives

$$w_i(\boldsymbol{x}) := \mathbb{P}_{\widehat{P}_0}(X_0 = \boldsymbol{z}_i \mid X_T = \boldsymbol{x}) = \frac{N_{\boldsymbol{z}_i}\boldsymbol{P}_T(\boldsymbol{z}_i,\boldsymbol{x})}{\sum_{j=1}^{D_n} N_{\boldsymbol{z}_j}\boldsymbol{P}_T(\boldsymbol{z}_j,\boldsymbol{x})}, \qquad i\in[D_n],$$

where $\boldsymbol{P}_T(\boldsymbol{z}_i,\boldsymbol{x}) = \prod_{\ell=1}^d \boldsymbol{P}_T^{\mathrm{coord}}(z_i^\ell, x^\ell)$.

Conditional on $I = i$ and $X_T = \boldsymbol{x}$, the state at time δ follows the forward bridge from $\boldsymbol{z}_i$ to $\boldsymbol{x}$. For each coordinate $\ell\in[d]$, define

$$B_{i,\ell}^{T,\delta}(a \mid x^\ell) := \frac{\boldsymbol{P}_\delta^{\mathrm{coord}}(z_i^\ell, a)\boldsymbol{P}_{T-\delta}^{\mathrm{coord}}(a, x^\ell)}{\boldsymbol{P}_T^{\mathrm{coord}}(z_i^\ell, x^\ell)}, \qquad a\in[K].$$

Since the forward process evolves independently across coordinates, the bridge law factorizes as

$$\mathbb{P}(X_\delta = \boldsymbol{y} \mid X_0 = \boldsymbol{z}_i, X_T = \boldsymbol{x}) = \prod_{\ell=1}^d B_{i,\ell}^{T,\delta}(y^\ell \mid x^\ell).$$

Thus, after sampling I, the K^d-state bridge can be sampled through d independent K-state distributions. The complete procedure is given in Algorithm 1.

Proposition D.1 (Correctness and computational cost)**.** *Conditional on the observations, Algorithm 1 returns an exact sample from Q_δ^{os}. If the distinct observations and their multiplicities have been stored, one execution requires $O(d(D_n + K))$ operations.*

Proof. Conditional on $I = i$ and $X_T = \boldsymbol{x}$, the bridge factorization gives

$$\mathbb{P}(Y = \boldsymbol{y} \mid I = i, X_T = \boldsymbol{x}) = \frac{\boldsymbol{P}_\delta(\boldsymbol{z}_i,\boldsymbol{y})\boldsymbol{P}_{T-\delta}(\boldsymbol{y},\boldsymbol{x})}{\boldsymbol{P}_T(\boldsymbol{z}_i,\boldsymbol{x})}.$$

Averaging over I yields

$$\mathbb{P}(Y = \boldsymbol{y} \mid X_T = \boldsymbol{x}) = \frac{\widehat{P}_\delta(\boldsymbol{y})\boldsymbol{P}_{T-\delta}(\boldsymbol{y},\boldsymbol{x})}{\widehat{P}_T(\boldsymbol{x})} = \overleftarrow{\boldsymbol{P}}_{T\to\delta,\widehat{P}}(\boldsymbol{x},\boldsymbol{y}).$$

Averaging over $X_T\sim\pi$ proves that $Y\sim Q_\delta^{\mathrm{os}}$.

Computing the D_n mixture weights requires $O(dD_n)$ operations. Conditional on I, evaluating and sampling the d coordinate bridges requires $O(dK)$ operations. Thus, the total cost is $O(d(D_n + K))$. □

Algorithm 1 Exact one-step empirical reverse sampling

Require: Times $0 \le \delta < T$, distinct observations $\{\boldsymbol{z}_i\}_{i=1}^{D_n}$, and counts $\{N_{\boldsymbol{z}_i}\}_{i=1}^{D_n}$
Ensure: A sample $Y \sim Q_\delta^{\mathrm{os}}$

Draw $X_T \sim \pi$
for $i = 1, \dots, D_n$ **do**
    Compute $\widetilde{w}_i \leftarrow N_{\boldsymbol{z}_i} \prod_{\ell=1}^{d} \boldsymbol{P}_T^{\mathrm{coord}}(z_i^\ell, X_T^\ell)$
end for
Draw $I \in [D_n]$ with $\mathbb{P}(I = i \mid X_T) = \dfrac{\widetilde{w}_i}{\sum_{j=1}^{D_n} \widetilde{w}_j}$
for $\ell = 1, \dots, d$ **do**
    Independently draw $Y^\ell \sim B_{I,\ell}^{T,\delta}(\cdot \mid X_T^\ell)$
end for
return $Y = (Y^1, \dots, Y^d)$

D.2 Statistical guarantees

We next show that the one-step sampler attains the TV and KL rates of Theorems 3 and 6 without constructing a fixed-grid score network.

Theorem D.2 (Statistical guarantees for the exact one-step sampler)**.** *For $0 \le \delta < T$, the following statements hold.*

(i) TV loss. *For every realization of the observations,*

$$d_{\mathrm{TV}}(P_0, Q_\delta^{\mathrm{os}}) \le d_{\mathrm{TV}}(P_0, \widehat{P}_0) + d\delta + de^{-T}.$$

Consequently, if $d\delta \lesssim \sqrt{\mathfrak{s}_n(P_0)/n}$ and $T \ge \log(d\sqrt{n/\mathfrak{s}_n(P_0)})$, then

$$\mathbb{E}_{P_0} d_{\mathrm{TV}}(P_0, Q_\delta^{\mathrm{os}}) \lesssim \sqrt{\frac{\mathfrak{s}_n(P_0)}{n}}.$$

(ii) KL loss. *Suppose additionally that $\delta > 0$. For every realization of the observations,*

$$D_{\mathrm{KL}}(P_0 \| Q_\delta^{\mathrm{os}}) \le D_{\mathrm{KL}}(P_0 \| \widehat{P}_\delta) + d \log\left(1 + (K-1)e^{-T}\right).$$

Consequently, setting $\delta = 1/n$ and choosing $T \ge \log(Kdn/\mathfrak{s}_n(P_0))$ gives

$$\mathbb{E}_{P_0} D_{\mathrm{KL}}(P_0 \| Q_{1/n}^{\mathrm{os}}) \lesssim \frac{\mathfrak{s}_n(P_0) \log\left(eK^d/\mathfrak{s}_n(P_0)\right) \log n}{n}.$$

Proof. Since $\widehat{P}_T \overleftarrow{\boldsymbol{P}}_{T \to \delta, \widehat{P}} = \widehat{P}_\delta$, contraction of total variation under the reverse kernel gives

$$d_{\mathrm{TV}}(\widehat{P}_\delta, Q_\delta^{\mathrm{os}}) \le d_{\mathrm{TV}}(\widehat{P}_T, \pi) \le de^{-T},$$

where the last inequality follows from the coordinatewise refresh coupling. The same coupling gives $d_{\mathrm{TV}}(\widehat{P}_0, \widehat{P}_\delta) \le d\delta$. Therefore,

$$d_{\mathrm{TV}}(P_0, Q_\delta^{\mathrm{os}}) \le d_{\mathrm{TV}}(P_0, \widehat{P}_0) + d\delta + de^{-T},$$

and the TV conclusion follows from Lemma A.6.

For the KL statement, the one-coordinate transition probabilities satisfy

$$\boldsymbol{P}_T^{\mathrm{coord}}(a, b) \le \frac{1 + (K-1)e^{-T}}{K}.$$

Hence

$$\widehat{P}_T(\boldsymbol{x}) \le \pi(\boldsymbol{x}) \left(1 + (K-1)e^{-T}\right)^d.$$

Using (D.1),

$$Q_\delta^{\mathrm{os}}(\boldsymbol{y}) \geq \left(1 + (K-1)e^{-T}\right)^{-d} \widehat{P}_\delta(\boldsymbol{y}),$$

which implies

$$D_{\mathrm{KL}}(P_0 \| Q_\delta^{\mathrm{os}}) \leq D_{\mathrm{KL}}(P_0 \| \widehat{P}_\delta) + d \log\left(1 + (K-1)e^{-T}\right).$$

For $\delta = 1/n$ and $T \geq \log(Kdn/\mathfrak{s}_n(P_0))$, the second term is at most $\mathfrak{s}_n(P_0)/n$. The KL conclusion therefore follows from Lemma A.11. □