DISCRETE DIFFUSION MODELS ARE MINIMAX OPTIMAL UNDER SPARSITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete diffusion models have emerged as a leading alternative to autoregressive models for generating text and other categorical data, yet their theoretical foundations remain less developed than those of their continuous counterparts. For continuous diffusion models, a prominent line of theory suggests that these models can adapt to low-dimensional structure in the data, thereby improving statistical efficiency and generalization. In this work, we investigate the discrete analogue of this phenomenon by studying the statistical rates of discrete diffusion models under sparsity assumptions. By analyzing a nonparametric estimator derived from a regularized score-entropy objective, we establish minimax-optimal rates in the sparse regime. The proposed regularization can be viewed as an adaptation of label smoothing techniques introduced for classification, and does not require the knowledge of the sparsity parameter. This shows that discrete diffusion models can adapt intrinsically to sparsity.

1 INTRODUCTION

Discrete diffusion models learn distributions on categorical data by reversing a Markov noising process (Austin et al., 2021; Lou et al., 2024). In language modeling, they provide an alternative to autoregressive sequence generation (Sahoo et al., 2024). We study their finite-sample estimation error in Kullback–Leibler (KL) divergence under two standard noising mechanisms: *uniform corruption*, which replaces tokens with uniformly sampled tokens, and *masked corruption*, which replaces them with an absorbing mask token. Our objective is to determine whether these mechanisms attain the optimal estimation rates when the data distribution has an unknown sparse support.

For continuous diffusion models, Oko et al. (2023) establish nearly minimax-optimal distribution estimation rates under smoothness assumptions. Related results show how score estimation and distribution recovery can exploit low-dimensional subspaces (Chen et al., 2023) and manifold structure (Azangulov et al., 2024). As discrete counterpart, we consider here support sparsity as structural assumption. For sequences of length d over a vocabulary of size K, the clean state space $\mathcal{X}_0$ has $M := K^d$ elements. An l-sparse distribution assigns positive probability to at most l sequences. We impose no geometric condition on these sequences, no independence assumption on their coordinates, and no lower bound on their positive probabilities. The relevant question is therefore whether diffusion-based estimation can adapt to l, rather than incur a statistical cost proportional to M.

Our starting point is to regularize the empirical score-entropy objective by its counterpart for uniformly distributed clean data. The resulting score is induced by the smoothed empirical distribution

$$\hat{\mu}_0^{(\gamma)} = \frac{n}{n+\gamma}\,\mu_0^{(n)} + \frac{\gamma}{n+\gamma}\,\mathrm{Unif}(\mathcal{X}_0), \qquad \gamma > 0, \tag{1}$$

where $\mu_0^{(n)}$ is the empirical law of the n training sequences. This distribution has full support, so it avoids the infinite forward KL loss that occurs when an estimator assigns zero probability to an unobserved sequence with positive data probability. The estimator in (1) does not depend on the corruption kernel; its forward marginals and associated scores do.

Smoothing discrete frequency estimates is classical. In particular, Mourtada (2026) analyzes an adaptive estimator of the form (1) and establishes sparse KL bounds and minimax lower bounds. Our contribution is to connect this estimation problem to regularized diffusion training and reverse sampling, and to determine how the smoothing level controls the resulting statistical error. For

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.

masked diffusion, this connection uses the conditional-distribution representation of the score established by Ou et al. (2025).

Our contributions can be summarized as follows:

- In the setting of masked diffusion, we derive a method for regularizing scores via a context-dependent label smoothing. We show that this smoothing is equivalent to regularizing with the score loss of the uniform distribution, which extends it beyond masked diffusion. The optimal conditional predictions that minimize the regularized loss produce the proper score induced by (1).
- We then establish expected and high-probability bounds for $\mathrm{KL}(\mu_{\mathrm{data}} \mid \hat{\mu}_0^{(\gamma)})$ in terms of the effective support and effective missing-support quantities of the data distribution, following the approach of Mourtada (2026). The bounds allow for data-dependent choices of γ, and the high-probability event holds simultaneously for every $\gamma \in [1, \min\{n, M\}]$. They therefore describe the effect of regularization beyond exact support sparsity. We also identify ranges of smoothing levels that attain the sparse minimax rates.
- We transfer these bounds to the terminal law $\hat{\nu}$ of the reverse sampler. For the masked diffusion initialized at the fully masked state, we prove the exact identity $\hat{\nu} = \hat{\mu}_0^{(\gamma)}$. For uniform diffusion, we control the finite-horizon initialization error. With the adaptive choice $\gamma = D_n$, where D_n is the number of distinct training sequences, both noising procedures attain a sparse minimax rate of order $l \log(eM/l)/n$ in expectation and high probability, without requiring knowledge of l or the support. Instead of depending on the size of the state space, which is exponentially large with respect to sequence length, the rate is instead governed by the sparsity l which can be far smaller in practice.

Relation to previous work. Recent analyses of discrete diffusion address several distinct statistical problems. Srikanth et al. (2026) derive sample-complexity guarantees for neural score estimators, accounting for statistical, approximation, optimization, and clipping errors. Wakasugi & Suzuki (2025) obtain state-size-independent bounds by exploiting smooth embeddings and graph structure; their assumptions differ from support sparsity. Cho & Wu (2026) derive nearly matching minimax bounds for score estimation under the score-entropy loss and apply their analysis to both uniform and masked diffusion. Their bounds depend on the cardinality of the diffusion state space and on density-ratio bounds; they do not quantify adaptation to the support size of the clean distribution.

Closer to our setting, Zhang et al. (2026) derive structure-dependent generalization bounds and compare the effects of sparsity under masked and uniform corruption. Such comparisons raise a further question: does the choice of corruption mechanism change the *minimax* risk over sparse data distributions? A difference between upper bounds for particular score estimators does not by itself establish a separation between the optimal risks. We address this question by analyzing a regularized score estimator and identifying the distribution generated by its reverse dynamics.

Notation. For a finite nonempty set $\mathcal{I}$, let $|\mathcal{I}|$ denote its cardinality, $\mathcal{P}(\mathcal{I})$ denote the set of probability measures on $\mathcal{I}$, $\mathrm{Unif}(\mathcal{I}) \in \mathcal{P}(\mathcal{I})$ be its uniform law, $\delta_x \in P(\mathcal{I})$ be the point mass at $x \in \mathcal{I}$. For $x \in \mathcal{X}$ and $A \subseteq [d]$, write $x_A := (x_j)_{j \in A}$ and $x_{-i} := x_{[d]\setminus\{i\}}$. For $\mu, \nu \in \mathcal{P}(\mathcal{I})$, define $\mathrm{KL}(\mu \mid \nu) := \sum_{x:\mu(x)>0} \mu(x) \log(\mu(x)/\nu(x))$, when $\mu \ll \nu$, and set $\mathrm{KL}(\mu \mid \nu) := +\infty$ otherwise. Given any positive integer K, we use the shorthand $[K] = \{1, ..., K\}$ and we use the notation $a \lesssim b$ to mean $a \leq Cb$ for a universal constant C.

2 PRELIMINARIES

We consider the problem of density estimation for a target distribution $\mu_{\mathrm{data}} \in \mathcal{P}(\mathcal{X}_0)$ on the finite sequence space, $\mathcal{X}_0 := \mathcal{V}^d$, $d \geqslant 1$. Without loss of generality, we take $\mathcal{V} := [K]$ with $K \geqslant 2$ leading to a cardinality $M := |\mathcal{X}_0| = K^d$ that grows exponentially with d.

Discrete diffusion models. The goal of discrete diffusion models is to learn a generative model that approximates μ_{data} with access to only a finite set of samples $S = (X_0^{(1)}, \ldots, X_0^{(n)}) \sim \mu_{\mathrm{data}}^{\otimes n}$. This is done by simulating the reversal of a Markov process that starts from the data distribution and

converges to a tractable limiting distribution. For convenience of exposition, we define this Markov process on a larger finite set $\mathcal{X} \supseteq \mathcal{X}_0$ and we identify μ_{data} with its extension by zero to $\mathcal{X}$.

The forward process is defined as a continuous-time Markov chain $(X_t)_{t\in[0,T]}$ on $\mathcal{X}$ with time horizon $T > 0$ and rate matrix $(Q_t)_{t\in[0,T]}$ (throughout this work we use the column-generator convention—details in Appendix A.1). We assume that $Q_t = \sigma_t Q_0$ where Q_0 is a fixed conservative generator and $\sigma : [0,\infty) \to [0,\infty)$ is integrable with $\bar{\sigma}_t := \int_0^t \sigma_u \, \mathrm{d}u$ satisfying $0 < \bar{\sigma}_t < \infty$ for all $t > 0$, and $\bar{\sigma}_t \to \infty$ as $t \to \infty$. For $0 \leqslant s \leqslant t \leqslant T$, $x, y \in \mathcal{X}$, we define the transition matrix $P_{s,t}(x,y) := \mathbb{P}(X_t = x \mid X_s = y)$ which satisfies,

$$P_{s,t} = \exp\big((\bar{\sigma}_t - \bar{\sigma}_s)Q_0\big) \, . \tag{2}$$

In particular, if $\mu_t(x) := \mathbb{P}(X_t = x)$, then μ_t solves the forward Kolmogorov equation,

$$\partial_t \mu_t(x) = Q_t \mu_t(x) = \sum_{y\in\mathcal{X}} Q_t(x,y)\mu_t(y) \, . \tag{3}$$

We initialize the process at the data distribution $\mu_0 = \mu_{\text{data}}$ so that, by definition, $\mu_t = P_{0,t}\mu_{\text{data}}$.

The reverse process. The time-reversed process, which we define by $\widetilde{X}_t := X_{T-t}$, $t \in [0,T]$, is also a continuous-time Markov chain whose rate matrix $\widetilde{Q}_t(x,y)$ satisfies the identity, $\mu_{T-t}(y)\widetilde{Q}_t(x,y) = \mu_{T-t}(x)Q_{T-t}(y,x)$ whenever $x \neq y$, $\mu_{T-t}(y) > 0$ (see, e.g., (Conforti & Léonard, 2022, Theorem 2.8)). This motivates the definition of the discrete score,

$$s^\star(x,y,t) := \begin{cases} \dfrac{\mu_t(x)}{\mu_t(y)}, & \mu_t(y) > 0, \\ 0, & \mu_t(y) = 0, \end{cases} \tag{4}$$

and hence, for $x \neq y$, the reverse generator can be written as $\widetilde{Q}_t(x,y) = s^\star(x,y,T-t)Q_{T-t}(y,x)$ with diagonal entries chosen so that the columns sum to zero. Equivalently, the reverse marginal law $\widetilde{\mu}_t$ solves

$$\partial_t \widetilde{\mu}_t = \widetilde{Q}_t \widetilde{\mu}_t, \qquad \widetilde{\mu}_0 = \mu_T. \tag{5}$$

With this initialization, we have that $\widetilde{\mu}_t = \mu_{T-t}$, and thus, exact sampling from the reverse process would recover the data distribution, $\widetilde{\mu}_T = \mu_{\text{data}}$.

In practice, the ratios $s^\star$ are unknown and must be estimated by some estimator score $s(x,y,t)$. We assume that this score is integrable in t and positive in the triple (x,y,t) where $Q_t(y,x) > 0$. We then define the associated generator $\hat{Q}_t(x,y)$ as $\tilde{Q}_t(x,y)$, but with $s(x,y,T-t)$ in place of $s^\star(x,y,T-t)$. We also do not have access to the initial distribution $\widetilde{\mu}_0 = \mu_T$ so we initialize at some tractable distribution $\mu_\infty \in \mathcal{P}(\mathcal{X})$ which is usually set to be a stationary distribution of the forward process. We define the sampling process on $[0,T)$ as the one which evolves according to

$$\partial_t \nu_t = \hat{Q}_t \nu_t, \qquad \nu_0 = \mu_\infty. \tag{6}$$

We will only use two choices of Q_0, corresponding to the uniform and masked noising processes, though we expect that our results may be adapted to other settings; see Appendix A.1 for details. For the uniform process, we set $\mathcal{X} = \mathcal{X}_0$ and consider a rate matrix with identical positive values on the off-diagonal, leading to a process that converges to the uniform measure. Therefore, in this case we set $\mu_\infty = \mathrm{Unif}(\mathcal{X})$. For the absorbing or masked process, we extend the space $\mathcal{X} = (\mathcal{V} \cup \{\mathfrak{m}\})^d$, where $\mathfrak{m} \notin \mathcal{V}$ is a mask token that the forward process converges to. In this case, we set $\mu_\infty = \delta_{(\mathfrak{m},\ldots,\mathfrak{m})}$.

Score-entropy loss. Following (Lou et al., 2024), in order to find the estimator s, we utilize the score-entropy loss, which is defined for the weights $w_t(x,y) \geqslant 0$ by,

$$\mathcal{L}_t(s) := \mathbb{E}_{y\sim\mu_t}\Big[\sum_{x\neq y} w_t(x,y)\Big(s(x,y,t) - s^\star(x,y,t)\log s(x,y,t) + \phi\big(s^\star(x,y,t)\big)\Big)\Big], \tag{7}$$

where $\phi(x) := x(\log x - 1)$, with $\phi(0) = 0$. Throughout, we take $w_t(x,y) = Q_t(y,x)$, following the literature, for example see (Zhang et al., 2026). In practice, we consider its empirical counterpart,

$$\hat{\mathcal{L}}_t(s) := \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{y\sim P_{0,t}(\cdot,X_0^{(i)})}\Big[\sum_{x\neq y} Q_t(y,x)\Big(s(x,y,t) - \frac{P_{0,t}(x,X_0^{(i)})}{P_{0,t}(y,X_0^{(i)})}\log s(x,y,t)\Big)\Big], \tag{8}$$

which is the score-entropy loss for the empirical law $\mu_0^{(n)} := n^{-1}\sum_{j=1}^n \delta_{X_0^{(j)}}$.

$l-$sparsity and lower bounds. In this work, we consider the class of l-sparse distributions. In particular, for a fixed integer l such that $1 \leqslant l \leqslant |\mathcal{X}_0|$, we define the class by,

$$\mathcal{P}_l(\mathcal{X}_0) := \{\mu \in \mathcal{P}(\mathcal{X}_0) : |\mathrm{supp}(\mu)| \leqslant l\}. \tag{9}$$

In this paper, we are interested in showing minimax rates for sparse density estimation and so we recall some lower bounds for the problem. This problem has been widely studied (Han et al., 2015; Kamath et al., 2015), though we focus on the following recent result of Mourtada (2026). Whenever $n, |\mathcal{X}| \geqslant 110$, $\delta \in (e^{-n}, e^{-2})$, $2 \leqslant l \leqslant \min\{n, |\mathcal{X}|/55\}$, for any estimator $\Phi : \mathcal{X}^n \to \mathcal{P}(\mathcal{X})$, there exists a distribution $P \in \mathcal{P}_l(\mathcal{X})$ such that,

$$\mathbb{P}_P\left(\mathrm{KL}(P|\hat{P}_n) \geqslant \frac{l\log(e|\mathcal{X}|/l) + \log(|\mathcal{X}|)\log(1/\delta)}{320\, n}\right) \geqslant \delta\,, \tag{10}$$

where $\hat{P}_n = \Phi(X_1, \ldots, X_n)$. From this, we immediately obtain bounds in expectation,

$$\mathbb{E}[\mathrm{KL}(P|\hat{P}_n)] \geqslant \frac{l\log(e|\mathcal{X}|/l)}{320e^2 n}.$$

Further results from (Mourtada, 2026) can be found in Appendix A.2

3 Statistical rates for masked diffusion

The goal of our analysis is to derive bounds for $\mathrm{KL}(\mu_{\mathrm{data}} \mid \hat{\nu})$ where we define $\hat{\nu} := \lim_{u\to T} \nu_u$, and ν_u is the solution of (6). In order to do this, we derive a particular choice of estimator $s(x, y, t)$, then we obtain upper bounds on $\mathrm{KL}(\mu_{\mathrm{data}} \mid \hat{\nu})$ as a function of a tuning parameter γ which we then compare with minimax lower bounds for the sparse distribution class. With this, we find two intervals for γ, one for the bound in expectation and one for the bound in high-probability, for which we reach the minimax optimal rates for the sparse distribution class.

Throughout this section, we denote by S our i.i.d. dataset, and denote by $\mathbb{E}_S$ and $\mathbb{P}_S$ the expectation and the probability with respect to the dataset. In the following, we begin with a problem of minimizing the empirical risk, which is a standard object of study in the training of discrete diffusion models. However, in the present setting, we cannot directly use the minimizer as it produces a score with infinite KL divergence. Thus, we look for non-parametric methods for regularizing the empirical problem.

3.1 Label smoothing for discrete scores

We first motivate our approach to regularizing the score by relating the score matching problem to classification. We define a *context* to be a triple $c = (A, v, i)$ where $A \subseteq [d]$, $v \in \mathcal{V}^A$ and $i \in [d] \setminus A$, and for a partially masked state y with masked position i we write $c(y, i) := (A(y), y_{A(y)}, i)$. In Ou et al. (2025), it was first shown that the score function $s^\star$ for the masked diffusion admits the factorization,

$$s^\star(x, y, t) = \kappa_t\, \psi^\star_{c(y,i)}(x_i), \qquad \kappa_t := \frac{\alpha_t}{1-\alpha_t}, \tag{11}$$

where $\psi^\star_c(a) := \mu_{\mathrm{data}}(X_i = a \mid X_A = v)$ is the conditional law of the clean token (see Proposition 2). Thus, the time dependence of the score can be obtained in closed-form, and the only quantity to learn is $\psi^\star_c$, which can be viewed as a *classifier* that predicts the token at position i from the unmasked tokens v. We thus parametrize the score estimator as $s(x, y, t) = \kappa_t\, \psi_{c(y,i)}(x_i)$ with $\psi_{c(y,i)} \in \mathcal{P}(\mathcal{V})$ which is also a parameterization used in practice, where the score network outputs a softmax over $\mathcal{V}$ at each masked position.

To take this analogue further, we can also describe the score-entropy loss via the cross-entropy. Let $n_c := |\{k : X^{(k)}_{0,A} = v\}|$, and let $\psi^{(n)}_c$ be the conditional distribution for the empirical data distribution, $\psi^{(n)}_c(a) := \mu^{(n)}_0(X_i = a \mid X_A = v)$. Substituting (11) into (8), it can be shown that

$$\int_0^T \hat{\mathcal{L}}_t(s)\, dt = \sum_c \omega_{|A|}\, \frac{n_c}{n}\, H\big(\psi^{(n)}_c, \psi_c\big) + \mathrm{const}, \tag{12}$$

where $H(p, q) := -\sum_a p(a)\log q(a)$ is the cross-entropy and the weights $\omega_k > 0$ depend only on k and T. The minimizer of (12) is the empirical conditional $\psi^{(n)}_c$ on every observed context, while contexts with $n_c = 0$ are left unconstrained.

By viewing ψ_c as a classifier trained with the cross-entropy loss, a natural choice of regularization is label smoothing (Szegedy et al., 2016). In classification, label smoothing adds the penalty $H(\mathrm{Unif}, \cdot)$ to the loss to reduce confidence in predictions and prevent overfitting. We thus add a label smoothing penalty to each classifier with a context-dependent parameter $\varepsilon_c \in [0, 1]$:

$$(1 - \varepsilon_c) H\big(\psi_c^{(n)}, \psi_c\big) + \varepsilon_c H\big(\mathrm{Unif}(\mathcal{V}), \psi_c\big). \tag{13}$$

By linearity of the cross-entropy in the first argument, we have that this is minimized by, $\hat{\psi}_c = (1 - \varepsilon_c)\, \psi_c^{(n)} + \varepsilon_c\, \mathrm{Unif}(\mathcal{V})$.

Note that for a generic choice of ε_c, it is not clear that the score function resulting from label smoothing will necessarily produce a *proper* score: that is, a score s that relates to a diffusion for a probability measure μ_s. In the following lemma, we show that there is a class of such choices of ε_c and that this class produces a smoothing of the empirical measure.

Lemma 1. *Suppose that for every $i \in [d]$, there exists $b_i \in \mathcal{V}$ such that $\mu_{\mathrm{data}}(X_i = b_i) = 0$, and that $\varepsilon_c = 1$ for every context $c = (A, v, i)$ with $n_c = 0$. Then, the score function $s(x, y, t) := \kappa_t \hat{\psi}_{c(y,i)}(x_i)$ is proper if and only if there exists $\gamma \in (0, \infty]$, not depending on c, such that*

$$\varepsilon_c = \frac{\gamma K^{-|A|}}{\gamma K^{-|A|} + n_c}, \qquad \textit{for every context } c = (A, v, i), \tag{14}$$

where $\varepsilon_c := 1$ if $\gamma = \infty$. In particular, s is the proper score induced by the terminal law, $\hat{\mu}_0^{(\gamma)}$.

We see that the smoothing is as in Szegedy et al. (2016), but where the level of smoothing ε_c decreases with the number of observations n_c of the context rather than being fixed. In particular, contexts that are observed frequently are barely smoothed, and contexts that do not appear in the sample are predicted uniformly. For this choice of ε_c, we find an elegant relationship between label smoothing and the score entropy loss relative to the uniform sequence distribution, $\mathcal{L}^U$. In particular, for the mixed loss $\hat{\mathcal{L}}_t^M(s) = (1 - \frac{\gamma}{n+\gamma})\, \hat{\mathcal{L}}_t(s) + \frac{\gamma}{n+\gamma}\, \mathcal{L}_t^U(s)$, we have,

$$\int_0^T \hat{\mathcal{L}}_t^M(s)\, dt = \sum_c \omega_{|A|}\, \tfrac{n_c + \gamma K^{-|A|}}{n+\gamma} \Big((1 - \varepsilon_c) H\big(\psi_c^{(n)}, \psi_c\big) + \varepsilon_c H\big(\mathrm{Unif}(\mathcal{V}), \psi_c\big) \Big) + \mathrm{const.}$$

For the statistical analysis pursued in this work, we take this class of proper label-smoothed score functions as the estimator which we analyze. In particular,

$$\hat{s}^{(\gamma)}(x, y, t) := \kappa_t \big((1 - \varepsilon_c)\, \psi_c^{(n)}(x_i) + \varepsilon_c\, \mathrm{Unif}(\mathcal{V}) \big), \qquad \varepsilon_c = \frac{\gamma K^{-|A|}}{\gamma K^{-|A|} + n_c}, \tag{15}$$

In the sections to follow, we analyze this estimator and show that for an appropriately tuned γ, we obtain minimax optimal rates in sparse settings.

3.2 Upper bounds

We now proceed with the statistical analysis of the estimator introduced above. To begin with, we introduce two fundamental data-dependent quantities that we borrow from (Mourtada, 2026) and use in our analysis.

Definition 1. *Given a probability distribution μ, let us define the effective support size $k_{\mathrm{eff}}(\mu, n)$ of μ at a sample size n as*

$$k_{\mathrm{eff}}(\mu, n) := \sum_{x \in \mathcal{X}_0} \min\{1, n\mu(x)\} = |\{x \in \mathcal{X}_0 : \mu(x) \geqslant 1/n\}| + \sum_{x : \mu(x) < 1/n} n\mu(x). \tag{16}$$

Similarly, we define the effective missing support size $k_{\mathrm{eff}}^o(\mu, n)$ of μ with sample size n as

$$k_{\mathrm{eff}}^o(\mu, n) := \sum_{x \in \mathcal{X}_0} \min\{\exp(1 - n\mu(x)), n\mu(x)\} = \sum_{x : \mu(x) \geqslant 1/n} \exp(1 - n\mu(x)) + \sum_{x : \mu(x) < 1/n} n\mu(x). \tag{17}$$

The quantity $k_{\text{eff}}(\mu, n)/n$ relates to error in the empirical measure due to low probability sequences that will likely be missing from the dataset. We note that $k_{\text{eff}}(\mu, n)$ is increasing with n while $k_{\text{eff}}(\mu, n)/n$ is decreasing. Also, $k_{\text{eff}}(\mu, 1) = 1$, $k_{\text{eff}}(\mu, n) \leqslant l$ if $\mu \in \mathcal{P}_l$, and $k_{\text{eff}}(\mu, n)$ converges to $k_{\text{eff}}(\mu, \infty) = \text{supp}(\mu)$ as $n \to \infty$.

We now provide two important lemmas which upper bound the KL between μ_{data} and $\hat{\mu}_0^{(\gamma)}$. Throughout this section, we require that γ is a measurable map from $\mathcal{X}_0^n$ to $[1, \min\{n, K^d\}]$. For the sake of brevity, we denote by $\hat{\mu}_0, k_{\text{eff}}, L_\gamma$ the objects $\hat{\mu}_0^{(\gamma(S))}, k_{\text{eff}}(\mu_{\text{data}}, n)$ and $\log \frac{eK^d}{\gamma}$, respectively.

Lemma 2. *If $n \geq 8$, then for every deterministic $\gamma_0 \in [1, \min\{n, K^d\}]$,*

$$\mathbb{E}_S \, \text{KL}(\mu_{\text{data}} | \hat{\mu}_0) \;\leq\; C \frac{k_{\text{eff}} + \mathbb{E}_S[\gamma] + k^o_{\text{eff}}(\mu_{\text{data}}, n/8)\, L_{\gamma_0}}{n} \;+\; \log(\gamma_0)\, \mathbb{P}_S\big(\gamma < \gamma_0\big), \tag{18}$$

for some universal constant $C > 0$.

Lemma 3. *If $n \geq 13$, $K^d \geq 3$ and $\delta \in (e^{-n/6}, e^{-2})$, there is an event A with $\mathbb{P}_S(A) \geq 1 - 12\delta$, defined in terms of S and independent of γ, on which, simultaneously for every $\gamma \in [1, \min\{n, K^d\}]$,*

$$\text{KL}\big(\mu_{\text{data}} \,\big|\, \hat{\mu}_0^{(\gamma)}\big) \;\leq\; C \frac{k_{\text{eff}} + \gamma + k^o_{\text{eff}}(\mu_{\text{data}}, n/112)\, L_\gamma + (1 + L_\gamma) \log \frac{1}{\delta} + \log \frac{1}{\delta} \log\log \frac{1}{\delta}}{n},$$

for some universal constant $C > 0$.

We note that if $\gamma_0 = 1$, the term $\log(\gamma_0)\mathbb{P}(\gamma < \gamma_0) = 0$, but the term L_{γ_0} takes the maximum value of $1 + d \log K$. Thus, we have a trade off, which will be optimized depending on the setting.

The lemmas above provide bounds, both in expectation and in high probability, on the KL divergence between μ_{data} and the smoothed empirical distribution $\hat{\mu}_0^{(\gamma)}$. However, to analyze the diffusion model, we must instead consider ν_t. We do this via the following lemma.

Lemma 4. *Let $\gamma > 0$, let $\hat{\mu}_0^{(\gamma)}$ be as in (1) and $\hat{s} = \hat{s}^{(\gamma)}$ as in (15). Then (6) with score $\hat{s}$ and initial law $\delta_{(m,\dots,m)}$ has a unique absolutely continuous solution $(\nu_u)_{u \in [0,T)}$, the limit $\hat{\nu} := \lim_{u \uparrow T} \nu_u$ exists, and $\hat{\nu} = \hat{\mu}_0^{(\gamma)}$, where $\hat{\mu}_0^{(\gamma)}$ is extended by zero to $\mathcal{X}$: $\hat{\nu}(x) = \hat{\mu}_0^{(\gamma)}(x)$ for $x \in \mathcal{X}_0$ and $\hat{\nu}(x) = 0$ for $x \in \mathcal{X} \setminus \mathcal{X}_0$. In particular, for every $\mu_{\text{data}} \in \mathcal{P}(\mathcal{X}_0)$,*

$$\text{KL}(\mu_{\text{data}} \,|\, \hat{\nu}) = \text{KL}(\mu_{\text{data}} \,|\, \hat{\mu}_0^{(\gamma)}).$$

Putting together the lemmas above, we obtain the following result.

Theorem 1. *Let γ be a measurable map from $\mathcal{X}_0^n$ to $[1, \min\{n, K^d\}]$, and denote by k_{eff}, L_γ the quantities $k_{\text{eff}}(\mu_{\text{data}}, n)$ and $\log \frac{eK^d}{\gamma}$, respectively. Then there exists a universal constant C such that the following hold:*

1. *If $n \geq 8$, then for every deterministic $\gamma_0 \in [1, \min\{n, K^d\}]$,*

$$\mathbb{E}_S \text{KL}(\mu_{\text{data}} \mid \hat{\nu}) \leqslant C \frac{k_{\text{eff}} + \mathbb{E}_S[\gamma] + k^o_{\text{eff}}(\mu_{\text{data}}, n/8) L_{\gamma_0}}{n} \;+\; \log(\gamma_0)\, \mathbb{P}_S\big(\gamma < \gamma_0\big). \tag{19}$$

2. *If $n \geq 13$, $K^d \geq 3$ and $\delta \in (e^{-n/6}, e^{-2})$, there is an event A with $\mathbb{P}_S(A) \geq 1 - 12\delta$ on which, simultaneously for every $\gamma \in [1, \min\{n, K^d\}]$,*

$$\text{KL}\big(\mu_{\text{data}} \,\big|\, \hat{\nu}\big) \leqslant C \frac{k_{\text{eff}} + \gamma + k^o_{\text{eff}}(\mu_{\text{data}}, \frac{n}{112}) L_\gamma}{n} + \frac{(1 + L_\gamma) \log \frac{1}{\delta} + \log \frac{1}{\delta} \log\log \frac{1}{\delta}}{n}. \tag{20}$$

The theorem above provides upper bounds of the error between μ_{data} and the limit of the sampling process (6). We highlight that this limit is not a mere theoretical object, but is the terminal law of an implementable algorithm based on the Gillespie method (Gillespie, 1977).

3.3 Sparse setting and minimax bounds

We now focus on the motivating setting where μ_{data} is sparse, i.e. $\mu_{\text{data}} \in \mathcal{P}_l(\mathcal{X}_0)$. We start by noticing that $k^o_{\text{eff}}(\mu_{\text{data}}, n), k_{\text{eff}}(\mu_{\text{data}}, n) \leqslant l$ for every n, and so, under no additional assumptions on γ, Theorem 1 produces the following corollary.

Corollary 1 (Upper bounds in the sparse setting). *Let us assume the hypothesis of Lemma 4. Moreover, let us consider the setting of Lemma 2 with $\mu_{\text{data}} \in \mathcal{P}_l(\mathcal{X}_0)$. Then, the bound becomes*

$$\mathbb{E}_S \mathrm{KL}(\mu_{\text{data}} \mid \hat{\nu}) \leqslant C \frac{l + \mathbb{E}[\gamma] + l L_{\gamma_0}}{n} + \log(\gamma_0) \mathbb{P}(\gamma < \gamma_0). \tag{21}$$

Similarly, in the setting of Lemma 3, with $\mu_{\text{data}} \in \mathcal{P}_l(\mathcal{X}_0)$ the bound becomes

$$\mathrm{KL}(\mu_{\text{data}} \mid \hat{\nu}) \leqslant C \frac{l + \gamma + l L_\gamma + (1 + L_\gamma) \log(1/\delta) + \log(1/\delta) \log\log(1/\delta)}{n}. \tag{22}$$

For deterministic γ, Corollary 1 produces a bound with rate $\frac{1}{n}(l \log \frac{eK^d}{\gamma} + \gamma)$. Thus, even with the generic choice of $\gamma = 1$, we obtain $O(\frac{ld \log(K)}{n})$ which no longer necessarily scales exponentially with dimension. However, the rates are still sub-optimal when l is large, and to achieve minimax optimal rates of $O(\frac{l \log(eK^d/l)}{n})$, we effectively require that $\gamma \sim l$. In the following corollary, we show that it is in fact sufficient for γ to be *concentrated* around l to obtain such rates. Throughout, we require that $n, K^d \geq 110$ and $2 \leq l \leq \min\{n, \lfloor K^d/55 \rfloor\}$; we write $\Lambda_l := \log(eK^d/l)$ and $\Gamma_\delta := l\Lambda_l + \log(K^d)\log(1/\delta)$, so that $\Gamma_1 = l\Lambda_l$. For a measurable $\gamma : \mathcal{X}_0^n \to [1, \min\{n, K^d\}]$ we denote by $\hat{\mu}_0 = \hat{\mu}_0^{(\gamma)}$ the estimator (1), by $\hat{s}$ the score, and by $\hat{\nu}$ the terminal law of (6) driven by $\hat{s}$.

Corollary 2. *Fix constants $0 < c_0 \leq 1 \leq C_0$. There exist a universal constant $c > 0$ and a constant $C < \infty$ depending only on (c_0, C_0) such that the following hold.*

(i) If γ takes values in $[\max\{1, c_0 l\},\ \min\{n, K^d, C_0\, l\Lambda_l\}]$, then for every $\mu_{\text{data}} \in \mathcal{P}_l(\mathcal{X}_0)$,

$$\mathbb{E}_S\, \mathrm{KL}(\mu_{\text{data}} \,|\, \hat{\nu}) \leq C\, \frac{l\Lambda_l}{n}.$$

(ii) Let $\delta \in (e^{-\min\{n/6, K^d\}}, e^{-2})$. If γ takes values in $[\max\{1, c_0 l\},\ \min\{n, K^d, C_0 \Gamma_\delta\}]$, then for every $\mu_{\text{data}} \in \mathcal{P}_l(\mathcal{X}_0)$, with probability at least $1 - 12\delta$,

$$\mathrm{KL}(\mu_{\text{data}} \,|\, \hat{\nu}) \leq C\, \frac{\Gamma_\delta}{n}.$$

We observe that the interval above depends on the sparsity parameter l, which is unknown in general. We find that a suitable estimator for such a quantity is given by,

$$D_n := |\mathrm{supp}\mu_0^{(n)}| = \#\{x \in \mathcal{X}_0 : N_0(x) > 0\}, \tag{23}$$

which denotes the number of different states observed in the sample at time $t = 0$. Using it, we can define an implementable, adaptive estimator, which reaches the optimal rates described above. The result is summarized in the Appendix B.3 (Corollary 4).

4 Statistical rates for other diffusions

We now extend outside of the masked setting to more general diffusions. For the masked diffusion, the factorization in (11) expresses the score as a classifier for the masked token multiplied by an explicit time-dependent factor, and it is this structure that motivated our use of label smoothing. For a general generator Q_0, no such factorization is available. We show, however, that the mixed score entropy loss $\mathcal{L}^M$ of Section 3.1 still admits an explicit minimizer, which smooths the empirical score in a manner analogous to label smoothing. We use this to extend our statistical guarantees to uniform corruption setting.

By linearity of the score entropy loss with respect to the data distribution, the mixed loss $\mathcal{L}^M$ can be expressed as a score entropy loss for the measure $\hat{\mu}_0^{(\gamma)}$ defined in (1). Thus, via the equivalence of the score entropy loss and the score matching loss, $\mathcal{L}^M$ is uniquely minimized by the score,

$$\hat{s}(x, y, t) := \frac{\hat{\mu}_t(x)}{\hat{\mu}_t(y)} = \frac{P_{0,t}\hat{\mu}_0^{(\gamma)}(x)}{P_{0,t}\hat{\mu}_0^{(\gamma)}(y)}. \tag{24}$$

Writing $\mu_t^{(n)} := P_{0,t}\mu_0^{(n)}$ for the noised empirical measure and $s^{\star,(n)}$ for its score, we have $\hat{\mu}_t = \frac{n}{n+\gamma}\mu_t^{(n)} + \frac{\gamma}{n+\gamma}\mu_t^U$, and dividing $\hat{\mu}_t(x)$ by $\hat{\mu}_t(y)$ gives

$$\hat{s}(x, y, t) = \big(1 - \varepsilon_t(y)\big)\, s^{\star,(n)}(x, y, t) + \varepsilon_t(y)\, s_U^\star(x, y, t), \quad \varepsilon_t(y) := \frac{\gamma\, \mu_t^U(y)}{\gamma\, \mu_t^U(y) + n\mu_t^{(n)}(y)}. \tag{25}$$

In the masked case, (25) recovers exactly the label smoothing of Section 3.1.

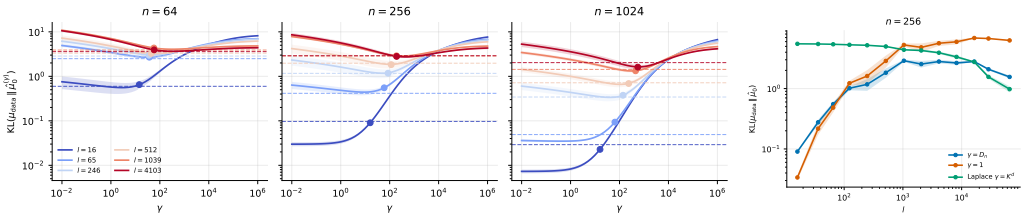

Figure 1: Label-smoothed score on restricted Markov chain dataset. Left: $\mathrm{KL}(\mu_{\mathrm{data}} \mid \hat{\nu})$ against γ, for $n \in \{64, 256, 1024\}$ for different values of l (blue to red). Dots mark the adaptive choice $\gamma = D_n$ and dashed lines are for the clipping estimator of Zhang et al. (2026) with tuned clipping parameter. Right: $\mathrm{KL}(\mu_{\mathrm{data}} \mid \hat{\nu})$ against l for different choices of γ demonstrating sparsity-adaptive behavior of small γ.

4.1 Uniform diffusion

We now consider the case of the uniform diffusion, for which we can obtain a neat expression of the mixed score-entropy loss as well as statistical rates for the limiting terminal distribution. In this case, $\mu_t^U = \mathrm{Unif}(\mathcal{X}_0)$ for every t, and so $s_U^\star \equiv 1$. Thus, the uniform score entropy loss takes the simple form,

$$\mathcal{L}_t^U(s) = \frac{1}{K^d} \sum_{y \in \mathcal{X}_0} \sum_{x \neq y} Q_t(y, x)\big(s(x, y, t) - \log s(x, y, t) - 1\big).$$

Furthermore, $\varepsilon_t(y) = \gamma/(\gamma + nK^d \mu_t^{(n)}(y))$ and the measure $\mu_t^{(n)}(y)$ has a closed-form expression:

$$\mu_t^{(n)}(y) = \frac{K^{-d}}{n} \sum_{k=1}^{n} \big(1 + (K-1)\alpha_t\big)^{d - H_k(y)} (1 - \alpha_t)^{H_k(y)} =: w_t(y),$$

where $H_k(y) := \#\{j \in [d] : y_j \neq X_{0,j}^{(k)}\}$ is the Hamming distance between y and the k-th sample. Turning to statistical rates, Lemmas 2 and 3 apply without modification, since they concern only the estimator $\hat{\mu}_0^{(\gamma)}$. The difference lies in the sampling step. For masked diffusion, the event that all tokens are masked at time T is independent of X_0, so that initializing the sampler at the fully masked state recovers $\hat{\mu}_0^{(\gamma)}$ exactly for every T (Lemma 4). Uniform diffusion admits no such state, and initializing the sampler at $\mathrm{Unif}(\mathcal{X}_0) \neq \hat{\mu}_T$ incurs an additional error, which is controlled in the following lemma.

Lemma 5 (Relative entropy of the sampler output)**.** *Let $\gamma > 0$, let $\hat{\mu}_0^{(\gamma)}$ be as in* (1) *and $\hat{s}$ as in* (24)*, and assume $\bar{\sigma}_t > 0$ for every $t \in (0, T]$. Let $(\nu_u)_u$ be the solution of* (6) *with score $\hat{s}$ and $\hat{\nu} := \lim_{u \uparrow T} \nu_u$ its terminal law. If $\bar{\sigma}_T \geq \log(dKn)$, then for every $\mu_{\mathrm{data}} \in \mathcal{P}(\mathcal{X}_0)$, the solution exists in all $[0, T]$, with $\hat{\nu} = \nu_T$ and $\mathrm{KL}(\mu_{\mathrm{data}} | \hat{\nu}) \leq \mathrm{KL}(\mu_{\mathrm{data}} | \hat{\mu}_0^{(\gamma)}) + \frac{1}{n}$.*

The initialization error thus contributes an additional term of order $1/n$. Since $k_{\mathrm{eff}} \geq 1$, this term is absorbed into k_{eff}/n up to a universal constant. As a result, uniform diffusion attains the same bounds as masked diffusion, at the cost of a logarithmic lower bound on the time horizon that is not required in the masked case. In the following theorem, we obtain end-to-end bounds for the limiting terminal law.

Theorem 2 (Uniform diffusion)**.** *Let γ be a measurable map from $\mathcal{X}_0^n$ to $[1, \min\{n, K^d\}]$ and consider the uniform diffusion with $\bar{\sigma}_T \geqslant \log(dKn)$ and terminal law $\hat{\nu} := \lim_{u \uparrow T} \nu_u$. Then, both statements of Theorem 1 hold for some different universal constant C.*

In particular, under the hypotheses of the theorem above, the minimax rates obtained for the masked setting in Corollary 1 and Corollary 2 transfer to the uniform setting also.

5 Experiments

We now provide some experiments that empirically evaluate our theoretical results. This also allows us to inspect the potential practical utility of two objects introduced in Section 3.1: the label-

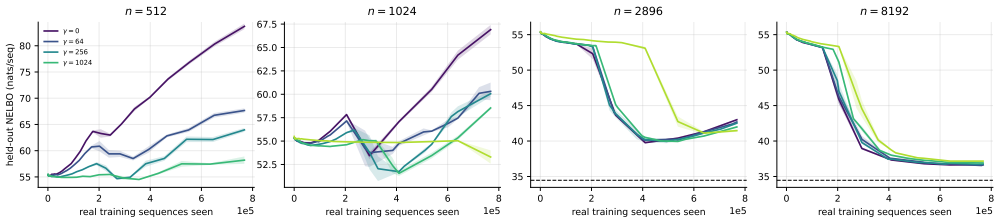

Figure 2: Transformer model trained with regularized score-entropy loss, $\mathcal{L}^M$ on synthetic Markov chain dataset. Training trajectories are plotted for different values of γ.

smoothed estimator, which requires no training, and the mixed loss $\mathcal{L}_t^M$, which we use to regularize a score network during training.

Exact label smoothing. We first analyze the effect of label smoothing by computing the closed form estimator $\hat{\psi}_c$ directly from dataset counts. As targets, we take the stationary laws of sparse Markov chains on $\mathcal{V} = [4]$ with $d = 8$, so that $K^d = 65{,}536$. Restricting the allowed transitions yields support sizes l ranging from 16 to 4096. The masked sampler returns exactly $\hat{\mu}_0^{(\gamma)}$ and so $\mathrm{KL}(\mu_{\mathrm{data}} \mid \nu)$ can be computed in closed form. Figure 1 shows that, for every l and n, the error forms a U-shape in γ and that the adaptive choice of $\gamma = D_n$ lands close to but not at its minimum, without any direct knowledge of l. We compare against the clipped empirical score of Zhang et al. (2026), which once tuned, obtains similar results to label-smoothing with $\gamma = D_n$ but performs worse than the optimal choice of γ.

Learning with the mixed loss. We now inspect the utility of the mixed score entropy loss $\mathcal{L}_t^M$ for training, which can be approximated by replacing each training sequence by a uniformly random one with probability $p = \gamma/(n + \gamma)$. We train the transformer of Zhang et al. (2026) on their synthetic Markov chain ($K = 100$, $d = 12$) and inspect train curves for different values of γ. Without regularization, the held-out NELBO rises sharply once the network starts memorizing (Figure 2). At $n = 1024$, the mixed loss removes this and improves the best held-out NELBO by 2.6 nats at $\gamma = 256$. For $n = 8192$ the network has enough data and the gain vanishes, but we find that regularization does not make results notably worse, likely due to how the amount of label-smoothing organically decreases with n.

Our results suggest that there may be some practical utility to utilizing label smoothing in the context of discrete diffusion. However, validating this will require a detailed ablation in more practical high dimensional settings, which we leave for future work.

6 Conclusion

In this work, we analyze the statistical efficiency of discrete diffusion models under sparsity. For the masked setting, we develop a regularized nonparametric estimator of the score function derived by treating it as a classifier and applying label smoothing. By regularizing the score in this fashion, we establish upper bounds on the KL divergence which achieve minimax optimal rates in the class of sparse distributions. We show that the same rates hold for uniform diffusion: the regularized score is again the exact score of (1), but the smoothing it induces has no interpretation as label smoothing of a classifier. Since the minimax lower bound is the same for both kernels, our results rule out a separation between masked and uniform diffusion at the level of the minimax risk over sparse distributions. Whether the gap observed in practice comes from approximation or optimization of the score, is a question we leave open. While our primary focus is providing theoretical justification for the success discrete diffusion, our work also suggests two practical methods of regularization for masked diffusions—label smoothing of the score and regularization via the uniform score-entropy loss—whose rigorous empirical evaluation we leave for future work.

AI use statement

In this work, we used generative AI tools for editing and assistance across the paper. We usually asked to the LLM to fix all the English mistakes, leaving as much as possible of the written text

already there. After that editing, anyway, we read everything and fixed little details. We also used AI to find references, that we checked immediately after. Finally, we used AI as a support for many proofs in the paper, especially to complete missing steps. Everything has been checked after. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work studies the statistical properties of an estimator for discrete diffusion models. The experiments use only synthetic data; no human subjects, personal data or real-world datasets are involved, and no data or models are released that could raise privacy or misuse concerns. We do not foresee ethical concerns beyond those common to research on generative modelling.

Reproducibility statement

The setting and assumptions are stated in Section 2 (continuous-time Markov chains in the column-generator convention, the uniform and masked kernels, the score-entropy loss) and Appendix A.1. Complete proofs of all results are given in Appendices B and C, with all numerical constants explicit. The experiments of Section 5 use synthetic data only: the sparse Markov chains and the closed-form label-smoothed estimator are fully specified in the text, and the transformer experiments follow the setup of Zhang et al. (2026), with the values of γ, n, K and d reported in Section 5 and Figures 2 and 1. Code to reproduce all figures will be released upon publication.

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 34*, 2021.

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv [stat.ML]*, September 2024.

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.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4672–4712. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chen23o.html.

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

Giovanni Conforti and Christian Léonard. Time reversal of markov processes with jumps under a finite entropy condition, 2022. URL https://arxiv.org/abs/2104.10948.

Devdatt P. Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *BRICS Report Series*, 3(25), Jan. 1996. doi: 10.7146/brics.v3i25.20006. URL https://tidsskrift.dk/brics/article/view/20006.

Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 12 1977. ISSN 0022-3654. doi: 10.1021/j100540a008. URL https://doi.org/10.1021/j100540a008.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ_1 loss, 2015. URL https://arxiv.org/abs/1411.1467.

Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In Peter Grünwald, Elad Hazan, and Satyen Kale (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1066–1100, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Kamath15.html.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning (ICML)*, 2024.

Jaouad Mourtada. Estimation of discrete distributions in relative entropy, and the deviations of the missing mass. *Mathematical Statistics and Learning*, 9(1):69–144, February 2026. ISSN 2520-2324. doi: 10.4171/msl/55. URL http://dx.doi.org/10.4171/msl/55.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning (ICML)*, 2023.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.03736.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.07524.

Aadithya Srikanth, Mudit Gaur, and Vaneet Aggarwal. Discrete state diffusion models: A sample complexity perspective, 2026. URL https://arxiv.org/abs/2510.10854.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

Shintaro Wakasugi and Taiji Suzuki. State size independent statistical error bound for discrete diffusion models. In D Belgrave, C Zhang, H Lin, R Pascanu, P Koniusz, M Ghassemi, and N Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 138908–138943. Curran Associates, Inc., 2025.

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

A BACKGROUND

A.1 BACKGROUND ON THE PROCESS

Throughout, as in Section 2, we use the column-generator convention: for $x \neq y$, $Q_t(x,y) \geqslant 0$ is the instantaneous rate of a jump from y to x, and $Q_t(y,y) = -\sum_{x \neq y} Q_t(x,y)$, so that the columns of Q_t sum to zero. Also, $P_{0,t}(x,y) = \mathbb{P}(X_t = x \mid X_0 = y)$, so that $\mu_t = P_{0,t}\mu_0$. Since $Q_t = \sigma_t Q_0$ with a fixed conservative Q_0, the generators commute across times and

$$P_{0,t} = \exp\Big(\int_0^t Q_u \, du\Big) = e^{\bar{\sigma}_t Q_0}, \qquad \bar{\sigma}_t = \int_0^t \sigma_u \, du, \tag{26}$$

with the semigroup property $P_{0,t_3} = e^{(\bar{\sigma}_{t_3} - \bar{\sigma}_{t_2})Q_0} P_{0,t_2}$ for $t_2 \leq t_3$. In both processes below, Q_0 acts independently on the d coordinates, $Q_0^{\mathrm{seq}} = \sum_{i=1}^d I^{\otimes(i-1)} \otimes Q_0^{\mathrm{tok}} \otimes I^{\otimes(d-i)}$, whence $P_{0,t}^{\mathrm{seq}} = (P_{0,t}^{\mathrm{tok}})^{\otimes d}$: the kernels factorize coordinatewise, and it suffices to compute the single-token propagator. Both token generators are of the form $Q_0^{\mathrm{tok}} = \Pi - I$ with Π a projection ($\Pi^2 = \Pi$), so

$$e^{\bar{\sigma}_t(\Pi - I)} = \Pi + e^{-\bar{\sigma}_t}(I - \Pi) = \alpha_t \, I + (1 - \alpha_t)\, \Pi, \qquad \alpha_t := e^{-\bar{\sigma}_t}. \tag{27}$$

Uniform process. On the token alphabet $\mathcal{V} = [K]$, take $\Pi := \frac{1}{K}\mathbf{1}\mathbf{1}^\top$ and $Q_0^{\mathrm{unif}} := \Pi - I_K$: every token jumps to each of the other $K-1$ values at rate $1/K$. By (27) and tensorization,

$$P_{0,t}^{\mathrm{unif}}(x, x_0) = \prod_{i=1}^d \Big[\alpha_t \mathbf{1}\{x_i = x_{0,i}\} + \tfrac{1-\alpha_t}{K}\Big].$$

The off-diagonal rates of Q_0^{seq} equal $1/K$, so $Q_0^{\max} = 1/K$, and the total exit rate of any state is $B_0 = d(K-1)/K \leq d$. The uniform law on $\mathcal{V}^d$ is stationary (each column of $\Pi - I$ sums to zero against the uniform vector).

Canonical absorbing process. On the token alphabet $\mathcal{V} \cup \{\mathfrak{m}\}$ (size $K+1$), take $\Pi := e_{\mathfrak{m}}\mathbf{1}^\top$ and $Q_0^{\mathrm{mask}} := \Pi - I_{K+1}$: every real token jumps to $\mathfrak{m}$ at unit rate, and $\mathfrak{m}$ is absorbing. By (27) and tensorization,

$$P_{0,t}^{\mathrm{mask}}(x, x_0) = \prod_{i=1}^d \Big[\alpha_t \mathbf{1}\{x_i = x_{0,i}\} + (1 - \alpha_t)\mathbf{1}\{x_i = \mathfrak{m}\}\Big],$$

and the point mass on the fully masked sequence is stationary. The total exit rate of a state y equals its number of unmasked coordinates, so $B_0 \leq d$. Finally, for a state y with masked coordinate $i \in M(y)$ and any $a \in \mathcal{V}$, the sequence-level rate from (y_{-i}, a) to y is the token-level rate $a \to \mathfrak{m}$, namely $Q_0\big(y, (y_{-i}, a)\big) = 1$.

A.2 USEFUL RESULTS

Lower bounds in the sparse setting.

Proposition 1 ((Mourtada, 2026), Section 4.1, Proposition 1)**.** *Let n be the number of samples $(X^{(1)}, \ldots, X^{(n)})$, and $n, |\mathcal{X}| \geqslant 2$. Let us assume $P \in \mathcal{P}_l(\mathcal{X})$. For any $1 \leqslant l \leqslant \min(n, |\mathcal{X}|/55)$ and any estimator $\hat{P}_n = \Phi(X^{(1)}, \ldots, X^{(n)})$, where Φ is an estimator map $\Phi : [|\mathcal{X}|]^n \to \mathcal{P}(\mathcal{X})$, there exists a distribution $P \in \mathcal{P}_l$ such that*

$$\mathbb{P}_P\left(\mathrm{KL}(P \mid \hat{P}_n) \geqslant \frac{l \log(e|\mathcal{X}|/l)}{300\, n}\right) \geqslant 1 - 3\exp\Big(-\frac{l}{35}\Big). \tag{28}$$

Corollary 3 ((Mourtada, 2026), Section 4.1, Corollary 1)**.** *Let us assume $P \in \mathcal{P}_l(\mathcal{X})$.Let n be the number of samples $(X^{(1)}, \ldots, X^{(n)})$, and $n, |\mathcal{X}| \geqslant 110$, $l \in \{2, \ldots, \min(n, |\mathcal{X}|/55)\}$ and $\delta \in (e^{-n}, e^{-2})$. For any estimator confidence-dependent $\Phi = \Phi_{n,\delta} : [|\mathcal{X}|]^n \to \mathcal{P}(\mathcal{X})$, there exists a distribution $P \in \mathcal{P}_l(\mathcal{X})$ such that, letting $\hat{P}_n = \Phi(X_1, \ldots, X_n)$, we have*

$$\mathbb{P}_P\left(\mathrm{KL}(P|\hat{P}_n) \geqslant \frac{l \log(e|\mathcal{X}|/l) + \log(|\mathcal{X}|)\log(1/\delta)}{320\, n}\right) \geqslant \delta\,. \tag{29}$$

Other results. We list here other results that are going to be useful.

In the following for $a \geq 0$ and $b > 0$ we set, with the convention $0 \log 0 := 0$,

$$D(a,b) := a \log \frac{a}{b} - a + b, \qquad \text{so that } D(0,b) = b.$$

Using ϕ, one has $D(a,b) = \phi(a) - \phi(b) - \phi'(b)(a-b)$, i.e. D is the Bregman divergence of ϕ. In particular $D(a,b) = b\,h(a/b)$ with $h(u) = u \log u - u + 1 \geq 0$, hence $D \geq 0$ with equality iff $a = b$; D is jointly convex and positively homogeneous, $D(ca, cb) = cD(a,b)$ for $c \geq 0$.

Lemma 6 (Relative entropy along two Markov flows)**.** *Let $\tau > 0$ and let $u \mapsto A_u, B_u$ measurable maps from $[0,\tau]$ to $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ with $\int_0^\tau M_u \mathrm{d}u < \infty$, where $M_u := \max_{x,y} \big(|A_u(x,y)| + |B_u(x,y)|\big)$, such that every A_u and B_u is a conservative generator in the column convention of Section 2. Let $(p_u)_{u\in[0,\tau]}$, $(q_u)_{u\in[0,\tau]}$ be the absolutely continuous solutions of $\dot p_u = A_u p_u$, $\dot q_u = B_u q_u$ with $p_0, q_0 \in \mathcal{P}(\mathcal{X})$ and $q_0 > 0$ on $\mathcal{X}$, and assume that the support S of p_u does not depend on $u \in [0,\tau]$. Then $q_u > 0$ on $\mathcal{X}$ for every u, the map $u \mapsto \mathrm{KL}(p_u|q_u)$ is absolutely continuous on $[0,\tau]$, and, writing $r_u := p_u/q_u$, for a.e. u*

$$\frac{d}{du}\mathrm{KL}(p_u|q_u) = \sum_{y\in\mathcal{X}} \sum_{x\neq y} \Big[A_u(x,y)\, p_u(y) \log \frac{r_u(x)}{r_u(y)} - B_u(x,y)\, q_u(y) \big(r_u(x) - r_u(y)\big) \Big], \quad (30)$$

where a term with $A_u(x,y)p_u(y) = 0$ is read as 0 (for a.e. u this covers every pair with $x \notin S$ or $y \notin S$). In particular, with $D(a,0) := +\infty$ for $a > 0$ and $D(0,0) := 0$, for a.e. u

$$\frac{d}{du}\mathrm{KL}(p_u|q_u) \leq \sum_{y\in\mathcal{X}} p_u(y) \sum_{x\neq y} D\big(A_u(x,y), B_u(x,y)\big). \quad (31)$$

Proof. Positivity of q. Solutions of Kolmogorov equations with bounded conservative generators stay in $\mathcal{P}(\mathcal{X})$, so $q_u \geq 0$ and, by the column convention, $\dot q_u(x) \geq -(|\mathcal{X}|-1)M_u\, q_u(x)$; Grönwall's lemma for integrable coefficients gives $q_u(x) \geq q_0(x) \exp\big(-(|\mathcal{X}|-1)\int_0^u M_z\,\mathrm{d}z\big) > 0$.

Absolute continuity. On S the functions $u \mapsto p_u(x)$ are continuous and positive, hence bounded away from 0 on $[0,\tau]$, and so are the $q_u(x)$. Thus $\mathrm{KL}(p_u|q_u) = \sum_{x\in S}[p_u(x)\log p_u(x) - p_u(x)\log q_u(x)]$ is a finite sum of products of absolutely continuous functions with smooth functions of absolutely continuous functions taking values in compact subsets of $(0,\infty)$; it is therefore absolutely continuous and can be differentiated term by term for a.e. u.

Two consequences of constant support. For a.e. u and every $x \notin S$, $0 = \dot p_u(x) = \sum_{y\neq x} A_u(x,y)p_u(y)$ is a sum of nonnegative terms, hence

$$A_u(x,y)\, p_u(y) = 0 \qquad \text{whenever } x \notin S, \text{ for a.e. } u, \quad (32)$$

and $\sum_{x\in S} \dot p_u(x) = \frac{d}{du}\sum_x p_u(x) = 0$.

Derivative. Omitting the index u, and using the two facts above and $r = 0$ off S,

$$\frac{d}{du}\mathrm{KL}(p|q) = \sum_{x\in S} \dot p(x) \log r(x) - \sum_{x\in S} r(x)\dot q(x).$$

By the column convention, $\dot p(x) = \sum_{y\neq x} A(x,y)p(y) - p(x)\sum_{y\neq x} A(y,x)$. Hence

$$\sum_{x\in S} \dot p(x)\log r(x) = \sum_{x\in S}\sum_{y\neq x} A(x,y)p(y)\log r(x) - \sum_{x\in S} p(x) \sum_{y\neq x} A(y,x)\log r(x).$$

In the second double sum only $y \in S$ contribute, by (32); exchanging the names of x and y there, and extending the first double sum to all x by (32) with the stated convention, we obtain $\sum_{y\in S} p(y)\sum_{x\neq y} A(x,y)[\log r(x) - \log r(y)]$. Similarly, using $r(x) = 0$ for $x \notin S$ and $q(y)r(y) = p(y)$,

$$\begin{aligned}\sum_{x\in S} r(x)\dot q(x) &= \sum_y \sum_{x\neq y} B(x,y)q(y)r(x) - \sum_y p(y) \sum_{x\neq y} B(x,y) \\ &= \sum_y \sum_{x\neq y} B(x,y)q(y)\big(r(x) - r(y)\big).\end{aligned}$$

This proves (30).

Bound. Split the right side of (30) according to $y \notin S$ or $y \in S$. If $y \notin S$ then $p(y) = r(y) = 0$ and the term is $-\sum_{x \neq y} B(x,y)q(y)r(x) \leq 0$. If $y \in S$, factor $q(y) = p(y)/r(y)$ and set $v := r(x)/r(y) \in [0, \infty)$: the term is $p(y)\sum_{x \neq y}[A \log v - Bv + B]$ with $A = A(x,y)$, $B = B(x,y)$. Now $A \log v - Bv + B \leq A \log \frac{A}{B} - A + B = D(A,B)$: if $A, B > 0$ and $v > 0$ this is $\log w \leq w - 1$ with $w = vB/A$, multiplied by A; if $A = 0$ both sides equal B; if $B = 0 < A$ the right side is $+\infty$; if $v = 0 < A$ the left side is $-\infty$ (a case excluded for a.e. u by (32) anyway). Summing gives (31). □

Lemma 7 (Support of the forward marginals)**.** *For both kernels and every $\rho \in \mathcal{P}(\mathcal{X}_0)$, the support of $\rho_t = P_{0,t}\rho$ is the same for all $t \in (0,T]$: it is $\mathcal{X}_0$ for the uniform kernel and $\{y \in \mathcal{X} : \rho(X_{A(y)} = y_{A(y)}) > 0\}$ for the masked kernel. In particular $\rho_t > 0$ on $\mathcal{X}$ for every $t \in (0,T]$ if $\rho > 0$ on $\mathcal{X}_0$, and on $[T_0, T]$, $T_0 > 0$, the positive entries of $P_{0,t}$ are bounded below by a positive constant.*

Proof. Immediate from the explicit kernels: $P^{\mathrm{unif}}_{0,t}(y, x_0) \geq ((1-\alpha_t)/K)^d > 0$ for all y, x_0, and $P^{\mathrm{mask}}_{0,t}(y, x_0) = \alpha_t^{|A(y)|}(1-\alpha_t)^{d-|A(y)|}\mathbf{1}\{x_{0,A(y)} = y_{A(y)}\}$, with $\alpha_t \in (0,1)$ for $t > 0$ and $\alpha_t \leq \alpha_{T_0}$, $1 - \alpha_t \geq 1 - \alpha_{T_0}$ on $[T_0, T]$. □

Lemma 8 (Data-processing inequality)**.** *Let $\mathcal{Y}, \mathcal{Z}$ be finite sets and let $\Pi \in \mathbb{R}^{|\mathcal{Z}| \times |\mathcal{Y}|}$ be a Markov kernel acting on measures, i.e. $\Pi(z,y) \geq 0$ and $\sum_z \Pi(z,y) = 1$ for every y. Then, for all $\rho, \bar{\rho} \in \mathcal{P}(\mathcal{Y})$,*

$$\mathrm{KL}(\Pi\rho \,|\, \Pi\bar{\rho}) \;\leq\; \mathrm{KL}(\rho \,|\, \bar{\rho}).$$

In particular:

- *(i) for every $0 \leq t_1 \leq t_2$, $\mathrm{KL}(\rho_{t_2}|\bar{\rho}_{t_2}) \leq \mathrm{KL}(\rho_{t_1}|\bar{\rho}_{t_1})$, where $\rho_t := P_{0,t}\rho$ and $\bar{\rho}_t := P_{0,t}\bar{\rho}$; that is, $t \mapsto \mathrm{KL}(\rho_t|\bar{\rho}_t)$ is nonincreasing, since $P_{0,t_2} = e^{(\bar{\sigma}_{t_2} - \bar{\sigma}_{t_1})Q_0}P_{0,t_1}$ and $e^{\tau Q_0}$ is a Markov kernel for every $\tau \geq 0$;*
- *(ii) for every $A \subseteq [d]$, $\mathrm{KL}(\rho_A|\bar{\rho}_A) \leq \mathrm{KL}(\rho|\bar{\rho})$, where $\rho_A(v) := \rho(X_A = v)$ is the marginal on the coordinates A, the projection $x \mapsto x_A$ being a (deterministic) Markov kernel.*

B Proofs of Section 3

B.1 Proofs of Section 3.1

Proof of Lemma 1. For $B \subseteq [d]$ and $w \in \mathcal{V}^B$ write $n_B(w) := \#\{k \leq n : X^{(k)}_{0,B} = w\}$; thus $n_c = n_A(v)$ depends on $c = (A, v, i)$ only through (A, v), and $\psi^{(n)}_c(a) = n_{A\cup\{i\}}(v, a)/n_A(v)$ when $n_c > 0$. By assumption $n_{A\cup\{i\}}(v, b_i) = 0$ for all A, v and $i \notin A$, hence

$$\hat{\psi}_c(b_i) = \varepsilon_c/K \qquad \text{for every context } c = (A, v, i), \tag{33}$$

since $\psi^{(n)}_c(b_i) = 0$ if $n_c > 0$, while $\hat{\psi}_c = \mathrm{Unif}(\mathcal{V})$ and $\varepsilon_c = 1$ if $n_c = 0$.

Step 0: reduction. Let $\bar{\mu} \in \mathcal{P}(\mathcal{X}_0)$ with $\bar{\mu} > 0$. By Proposition 2 applied to $\rho = \bar{\mu}$, for $t \in (0,T]$, $y \in \mathcal{X}$, $i \in M(y)$, $A = A(y)$ and $x = (y_{-i}, x_i)$,

$$s_{\bar{\mu}}(x, y, t) = \kappa_t\, \bar{\mu}(X_i = x_i \mid X_A = y_A),$$

and these are exactly the pairs (x, y) with $Q_t(y, x) > 0$ (Appendix A.1). Hence $s = \kappa_t\hat{\psi}_{c(y,i)}(x_i)$ is proper if and only if there is $\bar{\mu} > 0$ on $\mathcal{X}_0$ with

$$\bar{\mu}(X_i = a \mid X_A = v) = \hat{\psi}_c(a) \qquad \text{for all } c = (A, v, i),\ a \in \mathcal{V}. \tag{34}$$

Such a $\bar{\mu}$ is unique, since by the chain rule $\bar{\mu}(x) = \prod_{j=1}^d \hat{\psi}_{([j-1], x_{<j}, j)}(x_j)$.

If. Let $\gamma \in (0, \infty)$ and $\bar{\mu} := \hat{\mu}_0$, i.e. $\bar{\mu}(x) = (n_{[d]}(x) + \gamma K^{-d})/(n + \gamma) > 0$. We will often omit γ in the following, since it is fixed. Summing over the $K^{d-|A|}$ sequences x with $x_A = v$ gives

$\bar{\mu}(X_A = v) = (n_A(v) + \gamma K^{-|A|})/(n + \gamma)$, and likewise $\bar{\mu}(X_A = v, X_i = a) = (n_{A\cup\{i\}}(v, a) + \gamma K^{-|A|-1})/(n + \gamma)$. Dividing,

$$\bar{\mu}(X_i = a \mid X_A = v) = \frac{n_{A\cup\{i\}}(v, a) + \gamma K^{-|A|-1}}{n_A(v) + \gamma K^{-|A|}} = (1 - \varepsilon_c)\, \psi_c^{(n)}(a) + \frac{\varepsilon_c}{K}$$

with ε_c as in (14) (when $n_c = 0$ both sides equal $1/K$). Thus (34) holds, $s = s_{\bar{\mu}}$ is proper and $\mu_s = \hat{\mu}_0$. If $\gamma = \infty$, $\hat{\psi}_c = \mathrm{Unif}(\mathcal{V})$ for every c, which are the conditionals of $\mathrm{Unif}(\mathcal{X}_0)$.

Only if. Let $\bar{\mu} > 0$ satisfy (34). By (33)–(34) and positivity of $\bar{\mu}$,

$$\varepsilon_c = K\, \frac{\bar{\mu}(X_A = v, X_i = b_i)}{\bar{\mu}(X_A = v)} > 0 \qquad \text{for every } c = (A, v, i). \tag{35}$$

Let $A \subseteq [d]$ with $|A| \le d-2$, $v \in \mathcal{V}^A$, $i \neq j$ in $[d] \setminus A$ and $a, a' \in \mathcal{V}$. Computing $\bar{\mu}(X_A = v, X_i = a, X_j = a')$ in the two possible orders and dividing by $\bar{\mu}(X_A = v) > 0$,

$$\hat{\psi}_{(A,v,i)}(a)\, \hat{\psi}_{(A\cup\{i\},(v,a),j)}(a') = \hat{\psi}_{(A,v,j)}(a')\, \hat{\psi}_{(A\cup\{j\},(v,a'),i)}(a). \tag{36}$$

Step 1: ε_c depends only on (A, v). Take $a = b_i$, $a' = b_j$ in (36). The contexts $(A \cup \{i\}, (v, b_i), j)$ and $(A \cup \{j\}, (v, b_j), i)$ have zero count, so by assumption their $\hat{\psi}$ is $\mathrm{Unif}(\mathcal{V})$;then, the identity reads $\varepsilon_{(A,v,i)} K^{-2} = \varepsilon_{(A,v,j)} K^{-2}$. Hence $\varepsilon_{(A,v,i)} = \varepsilon_{(A,v,j)}$ whenever $|A| \le d-2$; if $|A| = d-1$ there is a single context with first two entries (A, v). We write $\varepsilon_{A,v}$.

Step 2: a one-parameter family. Take $a = b_i$ and a' arbitrary in (36). By (33), Step 1 and the fact that $(A \cup \{i\}, (v, b_i), j)$ has zero count, after multiplying by K,

$$\frac{\varepsilon_{A,v}}{K} = \hat{\psi}_{(A,v,j)}(a')\, \varepsilon_{A\cup\{j\},(v,a')} \qquad (|A| \le d-2,\ j \notin A,\ a' \in \mathcal{V}). \tag{37}$$

For (B, w) with $n_B(w) > 0$ and $|B| \le d-1$ set $g(B, w) := \varepsilon_{B,w}\, n_B(w)\, K^{|B|}/(1-\varepsilon_{B,w}) \in (0, \infty]$, with $g = \infty$ iff $\varepsilon_{B,w} = 1$ (recall $\varepsilon_{B,w} > 0$ by (35)). Let $|A| \le d-2$, $j \notin A$ and a' with $m' := n_{A\cup\{j\}}(v, a') > 0$; then $m := n_A(v) \ge m' > 0$ and, with $\varepsilon := \varepsilon_{A,v}$, $\varepsilon' := \varepsilon_{A\cup\{j\},(v,a')}$, we have $\hat{\psi}_{(A,v,j)}(a') = (1-\varepsilon)m'/m + \varepsilon/K$, so that (37) becomes $\varepsilon(1-\varepsilon')/K = (1-\varepsilon)\varepsilon' m'/m$, i.e.

$$\varepsilon\, m\, K^{|A|}(1 - \varepsilon') = (1 - \varepsilon)\, \varepsilon'\, m'\, K^{|A|+1}. \tag{38}$$

If $\varepsilon < 1$ and $\varepsilon' < 1$, (38) says $g(A, v) = g(A \cup \{j\}, (v, a'))$. If $\varepsilon = 1$ the left side vanishes, so $\varepsilon' = 1$ (as $m' > 0$); if $\varepsilon' = 1$ the right side vanishes, so $\varepsilon = 1$; in both cases $g(A, v) = g(A \cup \{j\}, (v, a')) = \infty$. Hence $g(A, v) = g(A \cup \{j\}, (v, a'))$ in every case.

Any (B, w) with $n_B(w) > 0$ and $|B| \le d-1$ is obtained from $(\emptyset, \emptyset)$ by adding the coordinates of B one at a time; every intermediate pair $(B', w_{B'})$ satisfies $n_{B'}(w_{B'}) \ge n_B(w) > 0$ and $|B'| \le d-2$ before the last step, so the previous paragraph applies at each step and $g(B, w) = g(\emptyset, \emptyset) =: \gamma \in (0, \infty]$ (note $n_\emptyset(\emptyset) = n$). Solving $g(B, w) = \gamma$ for $\varepsilon_{B,w}$ gives $\varepsilon_{B,w} = \gamma K^{-|B|}/(\gamma K^{-|B|} + n_B(w))$, which is (14) for all c with $n_c > 0$; for $n_c = 0$ we have $\varepsilon_c = 1$ by assumption, which is again (14). Finally, by the "if" part, $(\hat{\psi}_c)_c$ are the conditionals of $\hat{\mu}_0$, so $\bar{\mu} = \hat{\mu}_0$ by the uniqueness in Step 0. □

Lemma 9. *Consider the masked kernel. Let $s(x, y, t) = \kappa_t \psi_{c(y,i)}(x_i)$ with $\psi_c \in \mathcal{P}(\mathcal{V})$ for every context c, and for $k \in \{0, \dots, d-1\}$ set*

$$\omega_k := \int_0^T \sigma_t\, \alpha_t^{k+1} (1 - \alpha_t)^{d-k-1}\, dt \in (0, \infty).$$

Then

$$\int_0^T \hat{L}_t(s)\, dt = \sum_c \omega_{|A|}\, \frac{n_c}{n}\, H(\psi_c^{(n)}, \psi_c) + C_n,$$

$$\int_0^T L_t^U(s)\, dt = \sum_c \omega_{|A|} K^{-|A|}\, H(\mathrm{Unif}(\mathcal{V}), \psi_c) + C_U,$$

where $C_n, C_U \in \mathbb{R}$ do not depend on $(\psi_c)_c$. Consequently, for every $\gamma > 0$, with ε_c as in (14)*,*

$$\int_0^T \hat{L}_t^M(s)\,dt = \sum_c \omega_{|A|}\,\frac{n_c+\gamma K^{-|A|}}{n+\gamma}\Big[(1-\varepsilon_c)H(\psi_c^{(n)},\psi_c)+\varepsilon_c\,H(\mathrm{Unif}(\mathcal{V}),\psi_c)\Big]+\frac{nC_n+\gamma C_U}{n+\gamma},$$

which is minimized, uniquely, at $\psi_c = \hat{\psi}_c = (1-\varepsilon_c)\psi_c^{(n)} + \varepsilon_c \mathrm{Unif}(\mathcal{V})$ for every c. Likewise the first identity is minimized, uniquely on the contexts with $n_c > 0$, at $\psi_c = \psi_c^{(n)}$, while contexts with $n_c = 0$ are unconstrained.

Proof. Fix $t \in (0,T]$. For the masked kernel, $Q_t(y,x) > 0$ with $x \neq y$ if and only if $x = (y_{-i}, a)$ for some $i \in M(y)$ and $a \in \mathcal{V}$, and then $Q_t(y,x) = \sigma_t$ (Appendix A.1). Averaging (8) over the sample and using $\frac{1}{n}\sum_k P_{0,t}(y, X_0^{(k)}) = \mu_t^{(n)}(y)$, $\frac{1}{n}\sum_k P_{0,t}(x, X_0^{(k)}) = \mu_t^{(n)}(x)$, we get

$$\hat{L}_t(s) = \sigma_t \sum_{y\in\mathcal{X}} \sum_{i\in M(y)} \sum_{a\in\mathcal{V}} \Big[\mu_t^{(n)}(y)\,s(x,y,t) - \mu_t^{(n)}(x)\log s(x,y,t)\Big], \qquad x = (y_{-i}, a). \quad (39)$$

Fix y with $M(y) \neq \emptyset$, $i \in M(y)$, and let $c = (A, v, i)$ with $A = A(y)$, $v = y_A$. By the product form of $P_{0,t}^{\mathrm{mask}}$, $\mu_t^{(n)}(y) = \alpha_t^{|A|}(1-\alpha_t)^{d-|A|}\,n_c/n$ and $\mu_t^{(n)}(x) = \alpha_t^{|A|+1}(1-\alpha_t)^{d-|A|-1}\,n_c\,\psi_c^{(n)}(a)/n$ (both vanish if $n_c = 0$). Since $\kappa_t = \alpha_t/(1-\alpha_t)$ and $\sum_a \psi_c(a) = \sum_a \psi_c^{(n)}(a) = 1$, the inner sum over a in (39) equals

$$\alpha_t^{|A|+1}(1-\alpha_t)^{d-|A|-1}\,\frac{n_c}{n}\Big[1 - \log\kappa_t + H(\psi_c^{(n)},\psi_c)\Big].$$

As $(y,i) \mapsto c(y,i)$ is a bijection between $\{(y,i) : y \in \mathcal{X},\, i \in M(y)\}$ and the set of contexts,

$$\hat{L}_t(s) = \sum_c \sigma_t\,\alpha_t^{|A|+1}(1-\alpha_t)^{d-|A|-1}\,\frac{n_c}{n}\Big[1 - \log\kappa_t + H(\psi_c^{(n)},\psi_c)\Big],$$

and integrating in t gives the first identity with $C_n := \sum_c \frac{n_c}{n}\int_0^T \sigma_t\alpha_t^{|A|+1}(1-\alpha_t)^{d-|A|-1}(1 - \log\kappa_t)\,dt$. To check that ω_k and C_n are finite, substitute $u = \bar{\sigma}_t$ ($\sigma_t\,dt = du$, $\alpha_t = e^{-u}$): $\omega_k = \int_0^{\bar{\sigma}_T} e^{-(k+1)u}(1-e^{-u})^{d-k-1}du \in (0, \bar{\sigma}_T]$, and since $\log\kappa_t = -u - \log(1-e^{-u})$ with $|\log(1-e^{-u})| \le \log 2 + |\log u|$ (because $1-e^{-u} \ge u/2$ for $u \le 1$ and $1-e^{-u} \ge 1/2$ for $u \ge 1$), the integrand of C_n is bounded by $e^{-(k+1)u}(1+u+\log 2+|\log u|)$, which is integrable on $(0,\bar{\sigma}_T)$. The second identity is the same computation with $\mu_0^{(n)}$ replaced by $\mathrm{Unif}(\mathcal{X}_0)$, for which $\mathrm{Unif}(X_A = v) = K^{-|A|}$ and the conditional law of X_i given $X_A = v$ is $\mathrm{Unif}(\mathcal{V})$; the term $\phi(s_U^\star)$ in (7) does not depend on s and is absorbed in C_U. The identity for $\hat{L}_t^M = \frac{n}{n+\gamma}\hat{L}_t + \frac{\gamma}{n+\gamma}L_t^U$ follows since $\frac{n}{n+\gamma}\frac{n_c}{n} = \frac{n_c+\gamma K^{-|A|}}{n+\gamma}(1-\varepsilon_c)$ and $\frac{\gamma}{n+\gamma}K^{-|A|} = \frac{n_c+\gamma K^{-|A|}}{n+\gamma}\varepsilon_c$. Finally, all weights $\omega_{|A|}(n_c+\gamma K^{-|A|})/(n+\gamma)$ are positive, the cross-entropy is linear in its first argument, so $(1-\varepsilon)H(p,q)+\varepsilon H(u,q) = H((1-\varepsilon)p+\varepsilon u, q)$, and $H(p,q) = H(p,p) + \mathrm{KL}(p|q) \ge H(p,p)$ with equality iff $q = p$; this gives the two minimization statements. □

B.2 Proofs of Section 3.2

To begin, we recall here that a context c is a triple (A, v, i), where $A \subseteq [d]$, $v \in \mathcal{V}^A$ and $i \in [d] \setminus A$. In particular, given a certain state $y \in \mathcal{X}$, we will often call $c(y,i)$, or for brevity when it isn't ambiguous c, the triple $(A(y), y_{A(y)}, i)$, where $A(y)$ are the positions of unmasked tokens in y. We will denote also by $M(y) := [d] \setminus A(y)$, the position of the masked tokens in y. Given a position $i \in [d]$, a token $x_i \in \mathcal{V}$, and a state $y \in \mathcal{X}$ such that $y_i = \mathfrak{m}$, we can then define for a generic ρ the following conditional probability $\psi_c(a) := \frac{\rho(X_A = v, X_i = a)}{\rho(X_A = v)}$, when $\rho(X_A = v) > 0$.

Proposition 2 ((Ou et al., 2025), Theorem 1)**.** *Let $\rho \in \mathcal{P}(\mathcal{X}_0)$, $\rho_t := P_{0,t}^{\mathrm{mask}}\rho$. For every $t \in (0,T]$, every $y \in \mathcal{X}$ with $\rho_t(y) > 0$ (equivalently $\rho(X_{A(y)} = y_{A(y)}) > 0$), every $i \in M(y)$ and $a \in \mathcal{V}$,*

$$\frac{\rho_t(x)}{\rho_t(y)} = \kappa_t\,\rho(X_i = a \mid X_{A(y)} = y_{A(y)}), \qquad x := (y_{-i}, a), \quad \kappa_t = \frac{\alpha_t}{1-\alpha_t}.$$

Proof. By the product form of $P_{0,t}^{\mathrm{mask}}$ (Appendix A.1), $\rho_t(y) = \alpha_t^{|A|}(1-\alpha_t)^{d-|A|}\rho(X_A = y_A)$ with $A = A(y)$, and $\rho_t(x) = \alpha_t^{|A|+1}(1-\alpha_t)^{d-|A|-1}\rho(X_A = y_A, X_i = a)$. Divide. □

Let us now define an important object. Given a score s (measurable in t, positive where $Q_t(y,x) > 0$) define the *excess score entropy*

$$\mathcal{E}_t(s) := \sum_{y\in\mathcal{X}} \mu_t(y) \sum_{x\neq y} Q_t(y,x)\, D\big(s^\star(x,y,t),\, s(x,y,t)\big) \in [0,\infty], \tag{40}$$

with D as above (and $D(a,0) = +\infty$ for $a > 0$, $D(0,0) = 0$). Expanding D and using $\mu_t(y)s^\star(x,y,t) = \mu_t(x)$ whenever $Q_t(y,x) > 0$, one checks that $\mathcal{E}_t(s) = L_t(s)$, with L_t the score-entropy loss (7); in particular $L_t(s^\star) = 0$ and $L_t \geq 0$.

Throughout the section, many results hold also for the uniform case. For this reason we give already now their lemmas and proofs.

Lemma 10 (Lemma A: bounds on the uniform score)**.** *Consider the uniform kernel, $\gamma > 0$, and $\hat{s}$ the score for the uniform case. Then, for all $t \in [0,T]$ and $x, y \in \mathcal{X}_0$,*

$$M_\gamma^{-1} \leq \hat{s}(x,y,t) \leq M_\gamma, \qquad M_\gamma := \frac{K^d(n+\gamma)}{\gamma},$$

and $t \mapsto \hat{s}(x,y,t)$ is continuous. Moreover, for $0 < T_0 \leq T$, $s^\star(x,y,t) \leq K^d(1-\alpha_{T_0})^{-d}$ for all $t \in [T_0,T]$ and all $x, y \in \mathcal{X}_0$.

Proof. Q_0^{unif} is symmetric, so $P_{0,t} = e^{\bar{\sigma}_t Q_0}$ is symmetric with unit column sums, i.e. doubly stochastic. Hence $\hat{\mu}_t(x) = \sum_{x_0} P_{0,t}(x,x_0)\hat{\mu}_0^{(\gamma)}(x_0)$ is a convex combination of the values of $\hat{\mu}_0^{(\gamma)}$, so $\min \hat{\mu}_0^{(\gamma)} \leq \hat{\mu}_t(x) \leq \max \hat{\mu}_0^{(\gamma)}$, and $\min \hat{\mu}_0^{(\gamma)} = \gamma/(K^d(n+\gamma))$, $\max \hat{\mu}_0^{(\gamma)} \leq 1$. Dividing gives the two-sided bound; continuity in t is clear. For $s^\star$: $\mu_t(x) \leq 1$ and, by Appendix A.1, $\mu_t(y) \geq \min_{x_0} P_{0,t}(y,x_0) \geq ((1-\alpha_t)/K)^d \geq ((1-\alpha_{T_0})/K)^d$ on $[T_0,T]$. □

Lemma 11 (Lemma B: excess score entropy along the forward flow)**.** *Consider the uniform kernel, $\gamma > 0$, $\hat{\mu}_t := P_{0,t}\hat{\mu}_0^{(\gamma)}$, $\hat{s}$ as in* (24)*, and $0 < T_0 \leq T$. Then $t \mapsto \mathrm{KL}(\mu_t \,|\, \hat{\mu}_t)$ is absolutely continuous on $[T_0,T]$, for a.e. $t \in [T_0,T]$*

$$\frac{d}{dt}\mathrm{KL}(\mu_t \,|\, \hat{\mu}_t) = -\mathcal{E}_t(\hat{s}) = -L_t(\hat{s}),$$

and consequently

$$\int_{T_0}^{T} L_t(\hat{s})\mathrm{d}t = \mathrm{KL}(\mu_{T_0}|\hat{\mu}_{T_0}) - \mathrm{KL}(\mu_T|\hat{\mu}_T) \leq \mathrm{KL}(\mu_{T_0}|\hat{\mu}_{T_0}) \leq \mathrm{KL}(\mu_{\mathrm{data}} \,|\, \hat{\mu}_0^{(\gamma)}).$$

Proof. Apply Lemma 6 on $[T_0,T]$ with $A_t = B_t = Q_t = \sigma_t Q_0$ ($\int \sigma < \infty$), $p_t = \mu_t$, $q_t = \hat{\mu}_t$: $q_{T_0} = \hat{\mu}_{T_0} > 0$ because $\hat{\mu}_0^{(\gamma)} > 0$ (Lemma 7), and the support of μ_t is $\mathcal{X}_0$ for every $t > 0$ (Lemma 7), hence constant. Write $r_t := \mu_t/\hat{\mu}_t > 0$. By (30), for a.e. t,

$$\begin{aligned}\frac{d}{dt}\mathrm{KL}(\mu_t|\hat{\mu}_t) &= \sum_y \sum_{x\neq y} Q_t(x,y)\Big[\mu_t(y)\log\frac{r_t(x)}{r_t(y)} - \hat{\mu}_t(y)\big(r_t(x) - r_t(y)\big)\Big] \\ &= \sum_y \mu_t(y) \sum_{x\neq y} Q_t(x,y)\,\big(\log v_{xy} - v_{xy} + 1\big),\end{aligned}$$

where we used $\hat{\mu}_t(y) = \mu_t(y)/r_t(y)$ and set $v_{xy} := r_t(x)/r_t(y) = s^\star(x,y,t)/\hat{s}(x,y,t) \in (0,\infty)$. Exchange the names of x and y in the double sum: since $v_{yx} = 1/v_{xy}$ and $\mu_t(x) = \mu_t(y)s^\star(x,y,t) = \mu_t(y)\,\hat{s}(x,y,t)\,v_{xy}$, the summand becomes

$$\begin{aligned}\mu_t(y)\,Q_t(y,x)\,\hat{s}(x,y,t)\,v_{xy}\Big(-\log v_{xy} - \frac{1}{v_{xy}} + 1\Big) &= -\mu_t(y)\,Q_t(y,x)\,\hat{s}(x,y,t)\,h(v_{xy}) \\ &= -\mu_t(y)\,Q_t(y,x)\,D\big(s^\star(x,y,t), \hat{s}(x,y,t)\big),\end{aligned}$$

with $h(u) = u\log u - u + 1$ and $D(a,b) = b\,h(a/b)$ (Section 2). Summing gives $-\mathcal{E}_t(\hat{s})$. Integrate on $[T_0,T]$, use $\mathrm{KL} \geq 0$, and Lemma 8(i) for the last inequality. □

Lemma 12 (Transfer from score error to the sampler output)**.** *Let $0 < T_0 < T$, let $\mu_\infty \in \mathcal{P}(\mathcal{X})$ with $\mu_\infty > 0$ on $\mathcal{X}$, and let s be a score, measurable in t, with $\sup\{s(x,y,t) : x, y \in \mathcal{X},\ t \in [T_0, T]\} < \infty$ and $s(x,y,t) > 0$ whenever $s^\star(x,y,t)Q_t(y,x) > 0$. Let $(\nu_u)_{u\in[0,T-T_0]}$ be the solution of* (6) *with score s and initial law $\nu_0 = \mu_\infty$. Then*

$$\mathrm{KL}(\mu_{T_0} \,|\, \nu_{T-T_0}) \le \mathrm{KL}(\mu_T \,|\, \mu_\infty) + \int_{T_0}^{T} \mathcal{E}_t(s)\mathrm{d}t. \tag{41}$$

In particular, for the uniform kernel, $\mu_\infty = \mathrm{Unif}(\mathcal{X}_0)$ and $s = \hat{s}$ of (24)*, the hypotheses hold by Lemma A and* (41) *reads* $\mathrm{KL}(\mu_{T_0}|\nu_{T-T_0}) \le \mathrm{KL}(\mu_T|\mathrm{Unif}(\mathcal{X}_0)) + \int_{T_0}^{T} L_t(\hat{s})\mathrm{d}t$.

Proof. Set $\tau := T - T_0$ and, for $u \in [0,\tau]$ and $x \neq y$,

$$\tilde{Q}_u(x,y) := s^\star(x,y,T-u)\, Q_{T-u}(y,x), \qquad \hat{Q}_u(x,y) := s(x,y,T-u)\, Q_{T-u}(y,x),$$

with diagonals making the columns sum to zero. Both are measurable in u and $\int_0^\tau \max_{x,y}(|\tilde{Q}_u| + |\hat{Q}_u|)\mathrm{d}u < \infty$: indeed $|\hat{Q}_u(x,y)| \le \sup s \cdot \sigma_{T-u} \max|Q_0|$, and $s_t^\star$ is bounded on $[T_0, T]$ (masked kernel: $s^\star \le \kappa_t \le \kappa_{T_0}$ by Proposition 2; uniform kernel: Lemma A), with $\int_0^T \sigma < \infty$.

The two flows. Put $p_u := \mu_{T-u}$ and $q_u := \nu_u$. By definition q is the absolutely continuous solution of $\dot{q}_u = \hat{Q}_u q_u$, $q_0 = \mu_\infty > 0$. We claim $\dot{p}_u = \tilde{Q}_u p_u$ for a.e. u, with $p_0 = \mu_T$. For $x \in \mathcal{X}$,

$$(\tilde{Q}_u\mu_{T-u})(x) = \sum_{y\neq x} s^\star(x,y,T-u)Q_{T-u}(y,x)\mu_{T-u}(y) - \mu_{T-u}(x)\sum_{y\neq x} s^\star(y,x,T-u)Q_{T-u}(x,y).$$

Whenever $Q_t(y,x) > 0$ one has $s^\star(x,y,t)\mu_t(y) = \mu_t(x)$: if $\mu_t(y) > 0$ this is (4); if $\mu_t(y) = 0$ then $s^\star = 0$ and also $\mu_t(x) = 0$ (uniform kernel: $\mu_t > 0$, so this case does not occur; masked kernel: $Q_t(y,x) > 0$ means y is obtained from x by masking one coordinate, so $A(y) \subseteq A(x)$ and $\mu(X_{A(y)} = y_{A(y)}) = 0$ forces $\mu(X_{A(x)} = x_{A(x)}) = 0$). Applying this to both sums,

$$\begin{aligned}(\tilde{Q}_u\mu_{T-u})(x) &= \mu_{T-u}(x)\sum_{y\neq x} Q_{T-u}(y,x) - \sum_{y\neq x} Q_{T-u}(x,y)\mu_{T-u}(y)\\ &= -(Q_{T-u}\mu_{T-u})(x)\\ &= \frac{d}{du}\mu_{T-u}(x),\end{aligned}$$

by the column convention and (3). This is the time-reversal property behind (5).

Lemma 6. The support of $p_u = \mu_{T-u}$ does not depend on $u \in [0,\tau]$ by Lemma 7 (applied to $\rho = \mu_{\mathrm{data}}$ on $[T_0, T]$), and $q_0 > 0$. Moreover $\hat{Q}_u(x,y) > 0$ whenever $\tilde{Q}_u(x,y) > 0$ by the hypothesis on s, so $D(\tilde{Q}_u(x,y), \hat{Q}_u(x,y)) < \infty$ for every pair. Integrating (31) over $[0,\tau]$,

$$\mathrm{KL}(\mu_{T_0}|\nu_\tau) = \mathrm{KL}(p_\tau|q_\tau) \le \mathrm{KL}(p_0|q_0) + \int_0^\tau \sum_y \mu_{T-u}(y) \sum_{x\neq y} D\big(\tilde{Q}_u(x,y), \hat{Q}_u(x,y)\big)\mathrm{d}u.$$

Since $D(ca, cb) = cD(a,b)$ for $c \ge 0$ (with $D(0,0) = 0$), taking $c = Q_{T-u}(y,x)$ the inner sum equals $\sum_{x\neq y} Q_{T-u}(y,x)D(s^\star(x,y,T-u), s(x,y,T-u))$, and the substitution $t = T - u$ turns the integral into $\int_{T_0}^{T} \mathcal{E}_t(s)\mathrm{d}t$. This is (41). □

In order to prove Theorem 1, we need first the following lemmas.

Lemma 13 ((Mourtada, 2026), Section 6.1, Lemma 3)**.** *Let $\hat{P}_n = (\hat{p}_j)_j$ with $\hat{p}_j = \frac{N_j + \lambda}{n + \lambda K^d}$ for some $\lambda \in (0, n/K^d]$ that may depend on $X_1, \dots, X_n$, and let $\bar{p}_j := N_j/n$ be the empirical frequencies. Then*

$$\mathrm{KL}(P|\hat{P}_n) \le 6\sum_j \Big(\sqrt{\bar{p}_j} - \sqrt{p_j}\Big)^2 + \frac{7\lambda K^d}{n} + \sum_{j:\, p_j \ge 4\lambda/n} p_j \log\Big(\frac{2np_j}{\lambda}\Big)\mathbf{1}\Big(N_j \le \frac{np_j}{4}\Big).$$

Lemma 14 ((Mourtada, 2026),Lemma 4)**.** *We use the same notation as Lemma 13. For $n, K^d \geqslant 2$, and $\delta \in (0,1)$, with probability at least $1 - 2\delta$,*

$$\sum_{j\in\mathcal{X}_0} \Big(\sqrt{\bar{p}_j} - \sqrt{p_j}\Big)^2 < \frac{4k_{\mathrm{eff}}(\mu_{\mathrm{data}}, n) + 7\log\frac{1}{\delta}}{n}. \tag{42}$$

Lemma 15 ((Mourtada, 2026),Lemma 5)**.** *In the same setting as Lemma 13, for $n \geqslant 12, \lambda = 1, K^d \geqslant 2$ and $\delta \in (e^{-n/6}, e^{-2})$, with probability at least $1 - 2\delta$,*

$$R_\lambda := \sum_{j:\, p_j \geq 4\lambda/n} p_j \log\Big(\frac{2np_j}{\lambda}\Big)\mathbf{1}\Big(N_j \leq \frac{np_j}{4}\Big) \leqslant \frac{62000|\{x : p(x) \geqslant \frac{4}{n}\}| + 106000 \log\frac{1}{\delta}\log\log\frac{1}{\delta}}{n}. \tag{43}$$

Lemma 16 ((Mourtada, 2026), Theorem 6)**.** *In the same setting as Lemma 13, let*

$$U_n := \sum_x p(x)\mathbf{1}(N_x \leqslant np(x)/4). \tag{44}$$

For $n, K^d \geqslant 2$ and $\delta \in (0, e^{-1})$, with probability at least $1 - 8\delta$

$$U_n \leqslant \frac{336 k^o_{\mathrm{eff}}(\mu_{\mathrm{data}}, n/112) + 2500e\log\frac{1}{\delta}}{n}. \tag{45}$$

We define and recall now new objects that are going to be useful for the proof of the theorem. Given our sample $S = (X_0^{(1)}, \dots, X_0^{(n)})$, we call $\mu_0^{(n)} := \frac{1}{n}\sum_i \delta_{X_0^{(i)}}$ and $a_x := n\mu_{\mathrm{data}}(x)$. For $\gamma \in [1, \min\{n, K^d\}]$ we set $L_\gamma := \log(eK^d/\gamma) \geq 1$ and denote by $\hat{\mu}_0^{(\gamma)}$ the estimator (1) with parameter γ,

$$\hat{\mu}_0^{(\gamma)}(x) = \frac{N_0(x) + \gamma/K^d}{n + \gamma}, \qquad x \in \mathcal{X}_0.$$

The parameter γ is a measurable map $\gamma : \mathcal{X}_0^n \to [1, \min\{n, K^d\}]$, and $\hat{\mu} := \hat{\mu}_0^{(\gamma(S))}$. A deterministic γ is the special case of a constant map. Finally we use the three sample-dependent, parameter-free quantities

$$H_n := \sum_x \Big(\sqrt{\mu_0^{(n)}(x)} - \sqrt{\mu_{\mathrm{data}}(x)}\Big)^2, \qquad U_n := \sum_x \mu_{\mathrm{data}}(x)\,\mathbf{1}\big(N_0(x) \leq a_x/4\big),$$
$$V_n := \sum_{x:\, a_x \geq 1/2} \mu_{\mathrm{data}}(x)\log(2a_x)\,\mathbf{1}\big(N_0(x) \leq a_x/4\big),$$

which do not depend on γ and are nonnegative, since $\log(2a_x) \geq 0$ on the range of V_n. We observe that above U_n is the mass of the classes that empirically are observed less that one forth of the their true frequency.

Proof of Lemma 2. To start, we notice that, thanks to Lemma 13 we can immediately bound our object. In order to do that, we fix a sample S and a $\gamma \in [1, \min\{n, K^d\}]$. Taking $\lambda = \gamma/K^d$, we have that $\hat{p} = \hat{\mu}_0^{(\gamma)}$, and

$$\mathrm{KL}(\mu_{\mathrm{data}} \mid \hat{\mu}_0^{(\gamma)}) \leqslant 6H_n + \frac{7\gamma}{n} + R_\gamma, \tag{46}$$

where $R_\gamma := \sum_{x: a_x \geqslant 4\gamma/K^d} \mu_{\mathrm{data}}(x)\log\frac{2a_xK^d}{\gamma}\mathbf{1}(N_0(x) \leqslant a_x/4)$. Moreover, we can rewrite the logarithm as $\log(2a_x) + \log\frac{K^d}{\gamma}$. Then, $R_\gamma \leqslant \log\frac{K^d}{\gamma}U_n + V_n$, where $\log\frac{K^d}{\gamma} \geqslant 0$ because $\gamma \leqslant K^d$. We notice that above the range of the sums changed: in fact, in the case of V_n it includes only the terms for which $\log(2a_x) \geqslant 0$, i.e. $a_x \geqslant 1/2$, because the others are negative; in the other hand, the range in the case of U_n is the whole $\mathcal{X}_0$, so we are adding more positive terms. We then arrive to the following upper bound

$$\mathrm{KL}(\mu_{\mathrm{data}} \mid \hat{\mu}_0^{(\gamma)}) \leqslant 6H_n + \frac{7\gamma}{n} + V_n + \log\frac{K^d}{\gamma}U_n, \tag{47}$$

where H_n, V_n, U_n do not depend on γ.

We control now these three terms in expectation. To begin, we observe that when $\mu_{\mathrm{data}}(x) = 0$ then $\bar{p}(x) = 0$, so those x are negligible. For x such that $\mu_{\mathrm{data}}(x) > 0$, we use that $(\sqrt{a} - \sqrt{b})^2 = \frac{(a-b)^2}{(\sqrt{a}+\sqrt{b})^2} \leqslant \frac{(a-b)^2}{b}$ and that $(\sqrt{a} - \sqrt{b})^2 \leqslant a + b$. Moreover, $\mathbb{E}_S\mu_0^{(n)}(x) = \mu_{\mathrm{data}}(x)$ and, since our sample is i.i.d, $\mathbb{E}(\mu_0^{(n)}(x) - \mu_{\mathrm{data}}(x))^2 = \mu_{\mathrm{data}}(x)(1 - \mu_{\mathrm{data}}(x))/n$, thus

$$\mathbb{E}_S(6H_n) \leqslant 6\sum_x \min\{\frac{1}{n}, 2\mu_{\mathrm{data}}\} \leqslant \sum_x \frac{12}{n}\min\{1, a_x\} = \frac{12k_{\mathrm{eff}}}{n}. \tag{48}$$

Using again the fact that our sample is i.i.d, we observe immediately that $N_0(x) \sim \mathrm{Bin}(n, \mu_{\mathrm{data}}(x))$. Thanks to the multiplicative Chernoff bound, $\mathbb{P}(N_0(x) \leqslant (1-\theta)n\mu_{\mathrm{data}}(x)) \leqslant e^{-\theta^2 n\mu_{\mathrm{data}}(x)/2}$, so in particular $\mathbb{P}(N_0(x) \leqslant a_x/4) \leqslant e^{-9n\mu_{\mathrm{data}}(x)/32}$, where we took $\theta = \frac{3}{4}$. Hence,

$$\mathbb{E}_S U_n \leqslant \frac{1}{n}\sum_x a_x e^{-9a_x/32} \qquad \text{and} \qquad \mathbb{E}_S V_n \leqslant \frac{1}{n}\sum_{x:a_x \geqslant 1/2} a_x \log(2a_x) e^{-9a_x/32}. \tag{49}$$

We split the analysis of the terms inside the sums into two cases. If $a_x \leqslant 8$, then $a_x e^{-9a_x/32} \leqslant a_x$, since the exponent is negative, and $a_x \log(2a_x) e^{-9a_x/32} \leqslant a_x \log(16)$. On the other hand, if $a_x \geqslant 8$,

$$\begin{aligned} a_x e^{-9a_x/32} &= a_x e^{-5a_x/32} e^{-a_x/8} \\ &\leqslant e^{-a_x/8} \sup_{a_x>0}(a_x e^{-5a_x/32}) \\ &= e^{-a_x/8}\frac{32}{5e} \\ &\leqslant e^{-a_x/8} e \\ &\leqslant e^{1-a_x/8}. \end{aligned}$$

Moreover, $a_x \log(2a_x) e^{-9a_x/32} \leq \sqrt{2}\, a_x^{3/2} e^{-5a_x/32} e^{-a_x/8} \leq \sqrt{2}\, e^{-a_x/8} \sup_{a>0} a^{3/2} e^{-5a/32} = \sqrt{2}\left(\frac{48}{5e}\right)^{3/2} e^{-a_x/8} \leq 4e^{1-a_x/8}$, where above we used that $\log u \leqslant \sqrt{u}$ for $u > 0$ and $\sup_{a>0} a^p e^{-ca} = (\frac{p}{ce})^p$. If we call $m(a) := \min\{e^{1-a/8}, a/8\}$, we notice that we gave as upper bound always a multiple of $m(a_x)$. Noticing that $\sum_x m(a_x) = k^o_{\mathrm{eff}}(\mu_{\mathrm{data}}, n/8)$, we get

$$\mathbb{E}_S U_n \leqslant \frac{8k^o_{\mathrm{eff}}(\mu_{\mathrm{data}}, n/8)}{n} \qquad \text{and} \qquad \mathbb{E}_S V_n \leqslant \frac{23k^o_{\mathrm{eff}}(\mu_{\mathrm{data}}, n/8)}{n}. \tag{50}$$

We can now finish the proof. We just need to control the term $\log \frac{K^d}{\gamma} U_n$. We notice first of all that, since γ depends on S and the function $\gamma \mapsto \log \frac{K^d}{\gamma}$ is decreasing in $[\gamma_0, K^d]$, we have that if $\gamma \geqslant \gamma_0, \log \frac{K^d}{\gamma} \leqslant \log \frac{K^d}{\gamma_0} \leqslant \log \frac{eK^d}{\gamma_0} = L_{\gamma_0}$. On the other hand, if $\gamma < \gamma_0$, then

$$\begin{aligned} \log \frac{K^d}{\gamma} &= \log \frac{K^d}{\gamma_0} + \log \frac{\gamma_0}{\gamma} \\ &\leqslant \log \frac{K^d}{\gamma_0} + \log \gamma_0 \\ &\leqslant \log \frac{eK^d}{\gamma_0} + \log \gamma_0 \\ &= L_{\gamma_0} + \log \gamma_0, \end{aligned}$$

where we used that $\gamma \geqslant 1$. Then, to write both cases in a compact notation,

$$\log \frac{K^d}{\gamma} \leqslant L_{\gamma_0} + \log \gamma_0 \mathbf{1}(\gamma < \gamma_0). \tag{51}$$

Then, using $0 \leqslant U_n \leqslant 1$

$$\mathbb{E}_S[\log \frac{K^d}{\gamma} U_n] \leqslant L_{\gamma_0} \mathbb{E}_S U_n + \log \gamma_0 \mathbb{P}(\gamma < \gamma_0). \tag{52}$$

Putting all the results above together, we conclude. □

Proof of Lemma 3. To start, we recall the setting of the lemma above. Thanks to Lemma 13 we can immediately bound our object: fixing a sample S, a $\gamma \in [1, \min\{n, K^d\}]$, $\lambda = \gamma/K^d$, we have that $\hat{p} = \hat{\mu}_0^{(\gamma)}$, and

$$\mathrm{KL}(\mu_{\mathrm{data}} \mid \hat{\mu}_0^{(\gamma)}) \leqslant 6H_n + \frac{7\gamma}{n} + R_\gamma, \tag{53}$$

where $R_\gamma := \sum_{x:a_x \geqslant 4\gamma/K^d} \mu_{\text{data}}(x) \log \frac{2a_x K^d}{\gamma} \mathbf{1}(N_0(x) \leqslant a_x/4)$. Moreover, rewriting the logarithm, $R_\gamma \leqslant \log \frac{K^d}{\gamma} U_n + V_n$, where $\log \frac{K^d}{\gamma} \geqslant 0$ because $\gamma \leqslant K^d$. We then arrive to the following upper bound

$$\mathrm{KL}(\mu_{\text{data}} \mid \hat{\mu}_0^{(\gamma)}) \leqslant 6H_n + \frac{7\gamma}{n} + V_n + \log \frac{K^d}{\gamma} U_n,$$

where H_n, V_n, U_n do not depend on γ.

Let now A_1, A_2, A_3 be the events of Lemmas 14, 15 and 16 (the last one applies since $\delta < e^{-2} < e^{-1}$), of probabilities at least $1-2\delta$, $1-2\delta$, $1-8\delta$ respectively; each is defined through S only. On A_1,

$$6H_n \leq \frac{24 k_{\text{eff}} + 42 \log \frac{1}{\delta}}{n}. \tag{54}$$

For V_n, the terms with $\frac{1}{2} \leq a_x < 4$ satisfy $\log(2a_x) < \log 8$ and contribute at most $\log 8 \sum_{x:\, \mu_{\text{data}}(x) < 4/n} \mu_{\text{data}}(x)$, and by the second expression of k_{eff} in Definition 1,

$$\sum_{x:\, \mu_{\text{data}}(x) < 4/n} \mu_{\text{data}}(x) \leq \frac{1}{n} \Big[\sum_{x:\, \mu_{\text{data}}(x) < 1/n} a_x + 4 \, |\{x : \mu_{\text{data}}(x) \geq \frac{1}{n}\}| \Big] \leq \frac{4 k_{\text{eff}}}{n};$$

the terms with $a_x \geq 4$ are exactly those of R_{K^d}. Hence $V_n \leq \frac{4 \log 8 \, k_{\text{eff}}}{n} + R_{K^d}$ and, on A_2, using $|\{x : \mu_{\text{data}}(x) \geq 4/n\}| \leq k_{\text{eff}}$ and $4 \log 8 \leq 9$,

$$V_n \leq \frac{62009 \, k_{\text{eff}} + 106000 \log \frac{1}{\delta} \log\log \frac{1}{\delta}}{n}. \tag{55}$$

On A_3,

$$U_n \leq \frac{336 k^o_{\text{eff}}(\mu_{\text{data}}, n/112) + 2500 e \log \frac{1}{\delta}}{n} \tag{56}$$

with $2500e \leq 6796$. Set $A := A_1 \cap A_2 \cap A_3$, so $\mathbb{P}(A) \geq 1 - 12\delta$ by the union bound. On A, for every $\gamma \in [1, \min\{n, K^d\}]$, insert the three bounds and $\log \frac{K^d}{\gamma} \leq L_\gamma$ into (47):

$$\begin{aligned} \mathrm{KL}\big(\mu_{\text{data}} | \hat{\mu}_0^{(\gamma)}\big) \leq{}& \frac{(24 + 62009) k_{\text{eff}} + 7\gamma + 336 \, k^o_{\text{eff}}(\mu_{\text{data}}, n/112) L_\gamma}{n} \\ &+ \frac{(42 + 6796 L_\gamma) \log \frac{1}{\delta} + 106000 \log \frac{1}{\delta} \log\log \frac{1}{\delta}}{n}, \end{aligned}$$

which is the claim since $24 + 62009 \leq 63100$. The bound holds for all γ on the same event A, and A does not depend on the parameter; therefore, for any parameter γ, on A it holds. □

Proof of Lemma 4. For $y \in \mathcal{X}$ write $\hat{\mu}_0(y_{A(y)}) := \hat{\mu}_0^{(\gamma)}(X_{A(y)} = y_{A(y)})$, with $\hat{\mu}_0(y_\emptyset) := 1$; since $\hat{\mu}_0^{(\gamma)} > 0$ on $\mathcal{X}_0$, $\hat{\mu}_0(y_{A(y)}) > 0$ for every y. By the product form of $P^{\text{mask}}_{0,t}$,

$$\hat{\mu}_t(y) = \alpha_t^{|A(y)|} (1-\alpha_t)^{d - |A(y)|} \, \hat{\mu}_0(y_{A(y)}) > 0, \qquad t \in (0, T].$$

Thus, we can rewrite the estimator of (15) for x, y such that $Q_t(y, x) > 0$ has

$$\hat{s}(x, y, t) = \frac{\alpha_t}{1 - \alpha_t} \frac{\hat{\mu}_0(x_{A(x)})}{\hat{\mu}_0(y_{A(y)})} \qquad \text{with} \qquad x = (y_{-i}, x_i), i \in M(y). \tag{57}$$

We observe now that $\hat{Q}_u$ explodes when $u \to T$ since $\kappa_{T-u} \to \infty$ and that is why the we take the limit and we can not work on $[0, T]$. Also, recall that, for $x \neq y$, $\hat{Q}_u(y, x) = \hat{s}(y, x, T - u) \, Q_{T-u}(x, y)$, where $Q_{T-u}(x, y)$ is the forward rate from y to x. Since the forward process only masks tokens, $Q_{T-u}(x, y) \neq 0$ if and only if x is obtained from y by masking one unmasked coordinate, that is, if and only if $y = (x_{-i}, y_i)$ for some $i \in M(x)$ and $y_i \in \mathcal{V}$; in that case $Q_{T-u}(x, y) = \sigma_{T-u}$. Therefore, for such pairs

$$\hat{Q}_u(y, x) = \frac{\sigma_{T-u} \, \alpha_{T-u}}{1 - \alpha_{T-u}} \, \frac{\hat{\mu}_0(y_{A(y)})}{\hat{\mu}_0(x_{A(x)})}, \tag{58}$$

and $\hat{Q}_u(y,x) = 0$ for every other pair $x \neq y$. For simplicity, we are going to denote from now on $\lambda_u := \frac{\sigma_{T-u}\alpha_{T-u}}{1-\alpha_{T-u}}$. Then, equation (6) reads

$$\frac{d}{du}\nu_u(y) = \sum_{x\neq y} \hat{Q}_u(y,x)\nu_u(x) - \nu_u(y)\sum_{x\neq y}\hat{Q}_u(x,y), \qquad \text{with} \qquad \nu_0 = \delta_{\mathfrak{m}^d}. \tag{59}$$

Since for every T_0 such that $0 < T_0 < T$ and $u \in [0, T-T_0]$ we have $0 \leqslant \lambda_u \leqslant \frac{\sigma_{T-u}}{1-\alpha_{T_0}}$, that is integrable in u, we have that (59) is a linear system with locally integrable coefficients in $[0,T)$. By the Carathéodory existence and uniqueness theorem, since the system is linear with coefficients in $L^1_{\mathrm{loc}}([0,T))$, it has a unique absolutely continuous solution $(\nu_u)_{u\in[0,T)}$. We want to write it explicitly.

Since $u \in [0,T)$, the map $t \mapsto \log(1-\alpha_t)$ is absolutely continuous on $[T-u, T]$ with derivative $\frac{\sigma_t\alpha_t}{1-\alpha_t}$ a.e., so

$$\Lambda_u := \int_0^u \lambda_z dz = \int_{T-u}^T \frac{\sigma_t\alpha_t}{1-\alpha_t}dt = \log\frac{1-\alpha_T}{1-\alpha_{T-u}}. \tag{60}$$

Moreover, we define

$$F_u := 1 - e^{-\Lambda_u} = \frac{\alpha_{T-u}-\alpha_T}{1-\alpha_T} \in [0,1). \tag{61}$$

We focus on F. It is an absolutely continuous function on $[0,T)$ with $F_0 = 0, F_u \to 1$ as $u \to T$ due to the fact that $\alpha_{T-u} \to \alpha_0 = 1$, and $\frac{d}{dt}F_u = \lambda_u(1-F_u)$. Thanks to F_u, we can define

$$\tilde{\nu}_u(y) := F_u^{|A(y)|}(1-F_u)^{d-|A(y)|}\hat{\mu}_0(y_{A(y)}), \qquad y \in \mathcal{X}, u \in [0,T). \tag{62}$$

Above, we assumed the convention $0^0 = 1$.

We claim that $\tilde{\nu}_u$ is the explicit solution. First, we notice that the initial condition is $\delta_{\mathfrak{m}^d}$ since $F_0 = 0$ makes $\tilde{\nu}_0(y) = 0$ for all y such that $A(y) \neq \emptyset$. Fix now $y, k = |A(y)|$ and $A := A(y)$. By (59), we can write $\sum_{x\neq y}\hat{Q}_u(x,y) = \lambda_u \sum_{i\in M(y)}\sum_a \frac{\hat{\mu}_0(x_{A\cup i})}{\hat{\mu}_0(y_A)} = \lambda_u(d-k)$. Since the state x such that $\hat{Q}_u(y,x) \neq 0$ are exactly the ones that differs from y by only one token in a position $i \in A$, with a mask token in the case of x, thanks to (57) and the identity in (62), we have

$$\hat{Q}_u(y,x)\tilde{\nu}_u(x) = \lambda_u\frac{\hat{\mu}_0(y_A)}{\hat{\mu}_0(y_{A\setminus i})}F_u^{k-1}(1-F_u)^{d-k+1}\hat{\mu}_0(y_{A\setminus i}) = \lambda_u F_u^{k-1}(1-F_u)^{d-k+1}\hat{\mu}_0(y_A). \tag{63}$$

If we sum over $i \in A$, the right hand side of (59), at $\tilde{\nu}_u$ is equal to $\lambda_u\hat{\mu}_0(y_A)[kF_u^{k-1}(1-F_u)^{d-k+1} - (d-k)F_u^k(1-F_u)^{d-k}]$. Furthermore, for a.e. u we also have that

$$\begin{aligned}\frac{d}{du}[F_u^k(1-F_u)^{d-k}] &= [kF_u^{k-1}(1-F_u)^{d-k} - (d-k)F_u^k(1-F_u)^{d-k-1}]\frac{d}{du}F_u \\ &= \lambda_u[kF_u^{k-1}(1-F_u)^{d-k+1} - (d-k)F_u^k(1-F_u)^{d-k}].\end{aligned}$$

Above the terms with coefficient $k = 0$ or $d = k$ are absent. We can conclude then that the explicit solution for every $u \in [0,T)$ is $\tilde{\nu}_u$.

To finish this first point, for $x \in \mathcal{X}_0$, we have that $\nu_u(x) = F_u^d\hat{\mu}_0(x) \to \hat{\mu}_0(x)$, by continuity, and for $y \notin \mathcal{X}_0, 0 \leqslant \nu_u(y) \leqslant (1-F_u) \to 0$ as u goes to T. Thus, defining $\hat{\nu} := \lim_{u\to T}\nu_u$, we have that the limit exists and it is equal to $\hat{\mu}_0$. □

Proof of Theorem 1. Fix a sample S. Then $\gamma = \gamma(S) \in [1, \min\{n, K^d\}]$ is a number, $\hat{\mu}_0^{(\gamma(S))}$ is a strictly positive law on $\mathcal{X}_0$, and $\hat{s}$ of (15) is its score; Lemma 4 gives $\mathrm{KL}(\mu_{\mathrm{data}}|\hat{\nu}) = \mathrm{KL}(\mu_{\mathrm{data}}|\hat{\mu}_0^{(\gamma(S))})$ for this S. Item 1 follows by taking expectations and applying Lemma 2; item 2 follows by intersecting with the event A of Lemma 3, which does not depend on γ.

□

B.3 Proofs of Section 3.3

Throughout this subsection we write $k_{\mathrm{eff}} := k_{\mathrm{eff}}(\mu_{\mathrm{data}}, n)$, $L_\gamma := \log\frac{eK^d}{\gamma}$ for $\gamma \in [1, \min\{n, K^d\}]$ and $\Lambda_l := L_l = \log\frac{eK^d}{l}$. We use repeatedly the elementary facts

$$k^o_{\mathrm{eff}}(\mu_{\mathrm{data}}, m) \;\leq\; k_{\mathrm{eff}}(\mu_{\mathrm{data}}, m) \;\leq\; k_{\mathrm{eff}} \;\leq\; \min\{|\operatorname{supp}\mu_{\mathrm{data}}|,\, K^d\} \quad (m \leq n), \tag{64}$$

$u \mapsto u \log \frac{eK^d}{u}$ is nondecreasing on $(0, K^d]$, the first because $\min\{e^{1-a}, a\} \le \min\{1, a\} \le \min\{1, a'\}$ for $0 \le a \le a'$, the second because the derivative is $\log(K^d/u) \ge 0$.

Lemma 17 (Distinct observed states). *Let $s'_n := \mathbb{E}_S[D_n]$. Then:*

(a) $(1 - e^{-1})\, k_{\text{eff}} \le s'_n \le k_{\text{eff}}$;

(b) for every $\delta \in (0, 1)$, $\mathbb{P}_S\big(D_n \ge 2k_{\text{eff}} + 2\log\frac{1}{\delta}\big) \le \delta$;

(c) $\mathbb{P}_S\big(D_n \le 0.3\, k_{\text{eff}}\big) \le e^{-k_{\text{eff}}/21}$.

Proof. (a) Since $D_n = \sum_x \mathbf{1}\{N_0(x) > 0\}$ and $N_0(x) \sim \text{Bin}(n, \mu_{\text{data}}(x))$, we have $s'_n = \sum_x [1 - (1 - \mu_{\text{data}}(x))^n]$. For $p \in [0, 1]$, Bernoulli's inequality $(1-p)^n \ge 1 - np$ gives $1 - (1-p)^n \le \min\{1, np\}$; conversely $1 - (1-p)^n \ge 1 - e^{-np} \ge (1 - e^{-1}) \min\{1, np\}$, because $u \mapsto 1 - e^{-u}$ is nondecreasing and concave with value $1 - e^{-1}$ at $u = 1$. Summing over x and recalling $k_{\text{eff}} = \sum_x \min\{1, n\mu_{\text{data}}(x)\}$ gives (a).

(b)–(c) The indicators $\mathbf{1}\{N_0(x) > 0\}$, $x \in \mathcal{X}_0$, are negatively associated (Dubhashi & Ranjan, 1996), so $D_n = \sum_x \mathbf{1}\{N_0(x) > 0\}$ satisfies the Chernoff bounds of a sum of independent Bernoulli variables with the same means (Ben-Hamou et al., 2017): for every $s > s'_n$ and every $s < s'_n$ respectively,

$$\mathbb{P}(D_n \ge s) \le \exp\big(-D(s, s'_n)\big), \quad \mathbb{P}(D_n \le s) \le \exp\big(-D(s, s'_n)\big), \quad D(u, v) = u \log \frac{u}{v} - u + v.$$

(b) Let $L := \log(1/\delta)$ and $s := 2(s'_n + L)$. With $u := s/s'_n = 2 + 2\ell$, $\ell := L/s'_n$, we have $D(s, s'_n) = s'_n\, f(\ell)$, $f(\ell) := (2 + 2\ell)\log(2 + 2\ell) - (1 + 2\ell)$. Now $f(0) = 2\log 2 - 1 > 0$ and $f'(\ell) = 2\log(2 + 2\ell) \ge 2\log 2 > 1$, so $f(\ell) \ge \ell$ and $D(s, s'_n) \ge L$. Hence $\mathbb{P}(D_n \ge 2(s'_n + L)) \le \delta$, and since $s'_n \le k_{\text{eff}}$ by (a), $\{D_n \ge 2k_{\text{eff}} + 2L\} \subseteq \{D_n \ge 2(s'_n + L)\}$. (c) Take $s := s'_n/2$: $D(s'_n/2, s'_n) = \frac{1 - \log 2}{2} s'_n \ge \frac{(1-\log 2)(1-e^{-1})}{2} k_{\text{eff}} \ge k_{\text{eff}}/21$ by (a), and $s'_n/2 \ge \frac{1-e^{-1}}{2} k_{\text{eff}} \ge 0.3 k_{\text{eff}}$, so $\{D_n \le 0.3 k_{\text{eff}}\} \subseteq \{D_n \le s'_n/2\}$.

□

Proposition 3 ((Mourtada, 2026), Proposition 2). *Let $n, K^d \ge 2$. Then* $\mathbb{E}_S \,\text{KL}\big(\mu_{\text{data}} \,|\, \hat{\mu}_0^{(D_n)}\big) \le \frac{2.4\, k_{\text{eff}} + 2\, k^o_{\text{eff}}(\mu_{\text{data}}, n/2) \log\big(eK^d / k^o_{\text{eff}}(\mu_{\text{data}}, n/2)\big)}{n+1}$.

Corollary 4 (Adaptive smoothing). *Let $\hat{\mu}_0 = \hat{\mu}_0^{(D_n)}$ and let $k := k_{\text{eff}}$. Then the following hold.*

(i) If $n, K^d \ge 2$,

$$\mathbb{E}_S \,\text{KL}(\mu_{\text{data}} \,|\, \hat{\mu}_0) \le \frac{2.4\, k + 2\, k^o_{\text{eff}}(\mu_{\text{data}}, n/2) \log\big(eK^d / k^o_{\text{eff}}(\mu_{\text{data}}, n/2)\big)}{n+1}.$$

(ii) If $n \ge 13$, $K^d \ge 3$ and $\delta \in (e^{-n/6}, e^{-2})$, then with probability at least $1 - 14\delta$,

$$\text{KL}(\mu_{\text{data}} \,|\, \hat{\mu}_0) \le 130000\, \frac{k + k^o_{\text{eff}}(\mu_{\text{data}}, n/112) \log \frac{eK^d}{k} + \log(K^d) \log \frac{1}{\delta} + \log \frac{1}{\delta} \log\log \frac{1}{\delta}}{n}.$$

In particular, for $\mu_{\text{data}} \in \mathcal{P}_l(\mathcal{X}_0)$ with $2 \le l \le \min\{n, \lfloor K^d/55 \rfloor\}$, using $k^o_{\text{eff}}(\mu_{\text{data}}, m) \le k \le l$, the monotonicity of $u \mapsto u \log(eK^d/u)$ on $[1, \min\{n, K^d\}]$ and, for $\delta \ge e^{-K^d}$, $\log\log\frac{1}{\delta} \le \log(K^d)$, the right-hand sides are at most

$$C\, \frac{l \log \frac{eK^d}{l}}{n} \qquad \text{and} \qquad C\, \frac{\Gamma_\delta}{n},$$

simultaneously for every such l. Hence, under the assumptions of Corollary 2, the upper bounds in its items (i) and (ii) hold for $\hat{\nu}$ driven by $\hat{s} = \hat{s}^{(D_n)}$, with the same estimator and the same event for every l, and $\hat{\nu}$ is minimax optimal over $\mathcal{P}_l(\mathcal{X}_0)$ in both senses.

Proof of Corollary 4. (i) is Proposition 3.

(ii) Let A be the event of Lemma 3, $\mathbb{P}_S(A) \geq 1 - 12\delta$, and $B := \{D_n < 2k_{\text{eff}} + 2\log\frac{1}{\delta}\}$, $\mathbb{P}_S(B) \geq 1 - \delta$ by Lemma 17(b). Since $1 \leq D_n \leq \min\{n, K^d\}$, the parameter $\gamma = D_n$ is admissible in Lemma 3; on $A \cap B$, its explicit form with $\gamma = D_n$ and $7D_n \leq 14k_{\text{eff}} + 14\log\frac{1}{\delta}$ give

$$\mathrm{KL}(\mu_{\text{data}}|\hat{\mu}_0) \leq \frac{63114\, k_{\text{eff}} + 56\log\frac{1}{\delta} + \big(336\, k^o_{\text{eff}}(\mu_{\text{data}}, n/112) + 6796\log\frac{1}{\delta}\big) L_{D_n}}{n} + \frac{106000\log\frac{1}{\delta}\log\log\frac{1}{\delta}}{n}. \tag{65}$$

It remains to bound $L_{D_n} = \log(eK^d/D_n)$. Since $K^d \geq 3$, $\log K^d \geq 1$ and $\log(eK^d) = 1 + \log K^d \leq 2\log K^d$; since $D_n \geq 1$, $L_{D_n} \leq \log(eK^d) \leq 2\log K^d$ always.

Case $k_{\text{eff}} \geq 21\log\frac{1}{\delta}$. Let $E := \{D_n > 0.3\, k_{\text{eff}}\}$; by Lemma 17(c), $\mathbb{P}_S(E^c) \leq e^{-k_{\text{eff}}/21} \leq \delta$. On E,

$$L_{D_n} \leq \log\frac{eK^d}{0.3\, k_{\text{eff}}} = \log\frac{eK^d}{k_{\text{eff}}} + \log\frac{10}{3} \leq \log\frac{eK^d}{k_{\text{eff}}} + 1.21 \leq 2.21\log\frac{eK^d}{k_{\text{eff}}},$$

because $k_{\text{eff}} \leq K^d$ gives $\log\frac{eK^d}{k_{\text{eff}}} \geq 1$. Hence on $A \cap B \cap E$, of probability at least $1 - 14\delta$, $336\, k^o_{\text{eff}}(\mu_{\text{data}}, n/112) L_{D_n} \leq 743\, k^o_{\text{eff}}(\mu_{\text{data}}, n/112)\log\frac{eK^d}{k_{\text{eff}}}$ and $6796\log\frac{1}{\delta}\, L_{D_n} \leq 13592\log(K^d)\log\frac{1}{\delta}$.

Case $k_{\text{eff}} < 21\log\frac{1}{\delta}$. On $A \cap B$, of probability at least $1 - 13\delta$, using $k^o_{\text{eff}}(\mu_{\text{data}}, n/112) \leq k_{\text{eff}} < 21\log\frac{1}{\delta}$ and $L_{D_n} \leq 2\log K^d$, $336\, k^o_{\text{eff}}(\mu_{\text{data}}, n/112) L_{D_n} \leq 336 \cdot 21 \cdot 2\, \log(K^d)\log\frac{1}{\delta} = 14112\log(K^d)\log\frac{1}{\delta}$ and $6796\log\frac{1}{\delta}\, L_{D_n} \leq 13592\log(K^d)\log\frac{1}{\delta}$.

In both cases, inserting into (65) and using $56\log\frac{1}{\delta} \leq 56\log(K^d)\log\frac{1}{\delta}$, with probability at least $1 - 14\delta$,

$$\mathrm{KL}(\mu_{\text{data}}|\hat{\mu}_0) \leq \frac{63114\, k_{\text{eff}} + 743\, k^o_{\text{eff}}(\mu_{\text{data}}, n/112)\log\frac{eK^d}{k_{\text{eff}}} + 27760\log(K^d)\log\frac{1}{\delta}}{n} + \frac{106000\log\frac{1}{\delta}\log\log\frac{1}{\delta}}{n},$$

which implies the claim since $\max\{63114,\, 743,\, 27760,\, 106000\} \leq 130000$.

Sparse case. Let $\mu_{\text{data}} \in \mathcal{P}_l(\mathcal{X}_0)$ with $2 \leq l \leq \min\{n, \lfloor K^d/55 \rfloor\}$ and $\delta \geq e^{-K^d}$; note $\Lambda_l \geq 1$. By (64), $k^o_{\text{eff}}(\mu_{\text{data}}, m) \leq k_{\text{eff}} \leq l \leq K^d$ for $m \leq n$, hence $k_{\text{eff}} \leq l\Lambda_l$ and, by monotonicity, $k^o_{\text{eff}}(\mu_{\text{data}}, n/2)\log\frac{eK^d}{k^o_{\text{eff}}(\mu_{\text{data}}, n/2)} \leq l\Lambda_l$ and $k^o_{\text{eff}}(\mu_{\text{data}}, n/112)\log\frac{eK^d}{k_{\text{eff}}} \leq k_{\text{eff}}\log\frac{eK^d}{k_{\text{eff}}} \leq l\Lambda_l$. Therefore (i) gives $\mathbb{E}_S\mathrm{KL}(\mu_{\text{data}}|\hat{\mu}_0) \leq \frac{(2.4+2)\, l\Lambda_l}{n} = \frac{4.4\, l\Lambda_l}{n}$, and (ii) gives, with probability at least $1 - 14\delta$, since $\delta \geq e^{-K^d}$ implies $\log\log\frac{1}{\delta} \leq \log K^d$,

$$\mathrm{KL}(\mu_{\text{data}}|\hat{\mu}_0) \leq 130000\, \frac{2\, l\Lambda_l + 2\log(K^d)\log\frac{1}{\delta}}{n} = 260000\, \frac{\Gamma_\delta}{n}.$$

Neither $\hat{\mu}_0^{(D_n)}$ nor the events $A \cap B \cap E$, $A \cap B$ depend on l, so both bounds hold simultaneously for every l with $\mu_{\text{data}} \in \mathcal{P}_l(\mathcal{X}_0)$.

Transfer to the sampler. For every fixed sample S, $\hat{\mu}_0^{(D_n)}$ is a strictly positive law on $\mathcal{X}_0$ and $\hat{s}^{(D_n)}$ is its plug-in score, so Lemma 4, applied conditionally on S, gives $\mathrm{KL}(\mu_{\text{data}}|\hat{\nu}) = \mathrm{KL}(\mu_{\text{data}}|\hat{\mu}_0)$ for the masked kernel and $\mathrm{KL}(\mu_{\text{data}}|\hat{\nu}) \leq \mathrm{KL}(\mu_{\text{data}}|\hat{\mu}_0) + \frac{1}{n}$ for the uniform kernel (under $\bar{\sigma}_T \geq \log(dKn)$). Since $\frac{1}{n} \leq \frac{l\Lambda_l}{n} \leq \frac{\Gamma_\delta}{n}$, $\hat{\nu}$ satisfies the upper bounds of Corollary 2 with constants 5.4 and 260001, with the same event for every l. □

Proof of Corollary 2. Since the proof is the same for both kernels, we are going to explain it for both masked and uniform diffusion. Throughout $\mu_{\text{data}} \in \mathcal{P}_l(\mathcal{X}_0)$, so $k^o_{\text{eff}}(\mu_{\text{data}}, m) \leq k_{\text{eff}} \leq l$ for every $m \leq n$. Since $l \leq K^d/55$, $\Lambda_l \geq \log(55e) > 5$; since $K^d \geq 110$, $\Lambda_l \leq 1 + \log K^d \leq 2\log K^d$; and $1 \leq l \leq l\Lambda_l \leq \Gamma_\delta$.

Step 1: transfer to the sampler. Fix a sample S; $\hat{\mu}_0^{(\gamma(S))}$ is a strictly positive law on $\mathcal{X}_0$ and $\hat{s}$ is its score. Lemma 4 (masked kernel) gives $\mathrm{KL}(\mu_{\text{data}}|\hat{\nu}) = \mathrm{KL}(\mu_{\text{data}}|\hat{\mu}_0)$, and Lemma 5 (uniform

kernel, $\bar{\sigma}_T \geq \log(dKn)$) gives $\mathrm{KL}(\mu_{\mathrm{data}}|\hat{\nu}) \leq \mathrm{KL}(\mu_{\mathrm{data}}|\hat{\mu}_0) + 1/n$. As $1/n \leq l\Lambda_l/n \leq \Gamma_\delta/n$, it suffices to prove the upper bounds for $\hat{\mu}_0$ with constant $C-1$.

Step 2: a deterministic control of L_γ. Let $\gamma_0 := \max\{1, c_0 l\}$, a deterministic element of $[1, \min\{n, K^d\}]$ since $c_0 \leq 1$ and $l \leq \min\{n, K^d\}$; by assumption $\gamma \geq \gamma_0$ surely. If $c_0 l \geq 1$ then $L_{\gamma_0} = \log \frac{eK^d}{c_0 l} = \Lambda_l + \log \frac{1}{c_0}$; if $c_0 l < 1$ then $\gamma_0 = 1$ and $L_{\gamma_0} = \Lambda_l + \log l \leq \Lambda_l + \log \frac{1}{c_0}$. Since $u \mapsto L_u$ is nonincreasing,

$$L_\gamma \leq L_{\gamma_0} \leq \Lambda_l + \log \tfrac{1}{c_0} \leq c_1 \Lambda_l, \qquad c_1 := 1 + \frac{\log(1/c_0)}{\log(55e)},$$

where we used $\Lambda_l \geq \log(55e)$.

Step 3: upper bound in (i). Apply Lemma 2 (with $n \geq 110 \geq 8$) with this γ_0: $\mathbb{P}_S(\gamma < \gamma_0) = 0$, $\mathbb{E}_S[\gamma] \leq C_0 l \Lambda_l$, $k_{\mathrm{eff}} \leq l$, $k^o_{\mathrm{eff}}(\mu_{\mathrm{data}}, n/8) \leq l$. With the explicit constants of the proof of Lemma 2,

$$\mathbb{E}_S \mathrm{KL}(\mu_{\mathrm{data}}|\hat{\mu}_0) \leq \frac{12\, l + 7C_0\, l\Lambda_l + (23 + 8L_{\gamma_0})\, l}{n} \leq \frac{(12 + 7C_0 + 31c_1)\, l\Lambda_l}{n}.$$

With Step 1 this gives (i) with $C_1 := 13 + 7C_0 + 31c_1$.

Step 4: upper bound in (ii). Let A be the event of Lemma 3 (applicable: $n \geq 13$, $K^d \geq 3$, $\delta \in (e^{-n/6}, e^{-2})$), $\mathbb{P}_S(A) \geq 1 - 12\delta$, on which the bound holds simultaneously for every $\gamma \in [1, \min\{n, K^d\}]$, hence for the random γ. On A, with the explicit constants of the proof of Lemma 3 and $k_{\mathrm{eff}}, k^o_{\mathrm{eff}}(\mu_{\mathrm{data}}, n/112) \leq l \leq l\Lambda_l$, $\gamma \leq C_0\Gamma_\delta$, $L_\gamma \leq c_1\Lambda_l \leq 2c_1 \log K^d$, $\log(1/\delta) \leq \log(K^d)\log(1/\delta)$ and $\log\log(1/\delta) \leq \log K^d$ (as $\delta \geq e^{-K^d}$),

$$\begin{aligned}\mathrm{KL}(\mu_{\mathrm{data}}|\hat{\mu}_0) &\leq \frac{(63100 + 336c_1)\, l\Lambda_l + 7C_0\Gamma_\delta + (106042 + 13592c_1)\log(K^d)\log\frac{1}{\delta}}{n} \\ &\leq \frac{(169150 + 7C_0 + 13928c_1)\,\Gamma_\delta}{n}.\end{aligned}$$

With Step 1 this gives (ii) with $C_2 := 169151 + 7C_0 + 13928c_1$. Set $C := \max\{C_1, C_2\}$, which depends only on (c_0, C_0).

Step 5: lower bounds. For completeness, we also prove that the upper bounds of (i) and (ii) match the minimax lower bounds up to universal constants. Let $S \mapsto \nu_S \in \mathcal{P}(\mathcal{X})$ be measurable. Define $\tilde{\nu}_S := \nu_S(\cdot \mid \mathcal{X}_0)$ if $\nu_S(\mathcal{X}_0) > 0$ and $\tilde{\nu}_S :=$ a fixed element of $\mathcal{P}(\mathcal{X}_0)$ otherwise (for the uniform kernel $\mathcal{X} = \mathcal{X}_0$ and $\tilde{\nu}_S = \nu_S$). For every $\mu \in \mathcal{P}(\mathcal{X}_0)$, $\mathrm{KL}(\mu|\nu_S) \geq \mathrm{KL}(\mu|\tilde{\nu}_S)$: if $\nu_S(\mathcal{X}_0) = 0$ the left side is $+\infty$, otherwise $\tilde{\nu}_S \geq \nu_S$ on $\mathcal{X}_0$. The map $S \mapsto \tilde{\nu}_S$ is an estimator $\mathcal{X}_0^n \to \mathcal{P}(\mathcal{X}_0)$, so Corollary 3 applies with $|\mathcal{X}| = K^d$; its hypotheses $n, K^d \geq 110$, $l \in \{2, \ldots, \min(n, \lfloor K^d/55 \rfloor)\}$ are ours, and Γ_δ is exactly the numerator in (29).

(i) Take $\delta_0 := e^{-3} \in (e^{-n}, e^{-2})$ (as $n \geq 110$). Corollary 3 gives $\mu_{\mathrm{data}} \in \mathcal{P}_l(\mathcal{X}_0)$ with $\mathbb{P}_S\big(\mathrm{KL}(\mu_{\mathrm{data}}|\tilde{\nu}_S) \geq \frac{l\Lambda_l + 3\log K^d}{320n}\big) \geq e^{-3}$, whence, since $Z \geq a\mathbf{1}\{Z \geq a\}$ for $Z \geq 0$,

$$\mathbb{E}_S \mathrm{KL}(\mu_{\mathrm{data}}|\nu_S) \geq \mathbb{E}_S \mathrm{KL}(\mu_{\mathrm{data}}|\tilde{\nu}_S) \geq \frac{e^{-3}}{320}\,\frac{l\Lambda_l}{n}.$$

(ii) For $\delta \in (e^{-\min\{n/6, K^d\}}, e^{-2}) \subseteq (e^{-n}, e^{-2})$, Corollary 3 gives $\mu_{\mathrm{data}} \in \mathcal{P}_l(\mathcal{X}_0)$ with $\mathbb{P}_S\big(\mathrm{KL}(\mu_{\mathrm{data}}|\tilde{\nu}_S) \geq \frac{\Gamma_\delta}{320n}\big) \geq \delta$, and $\mathrm{KL}(\mu_{\mathrm{data}}|\nu_S) \geq \mathrm{KL}(\mu_{\mathrm{data}}|\tilde{\nu}_S)$.

Both lower bounds hold with the universal constant $c := e^{-3}/320$. □

C Proofs of Section 4

Lemma 18 (Lemma C). *Let $\gamma > 0$, $\hat{\mu}_t := P_{0,t}\hat{\mu}_0^{(\gamma)}$ and $\hat{s}(x, y, t) := \hat{\mu}_t(x)/\hat{\mu}_t(y)$ as in* (24)*. For every $t \in (0, T]$ and every score s positive where $Q_t(y, x) > 0$,*

$$\hat{L}_t^M(s) = \sum_{y \in \mathcal{X}} \hat{\mu}_t(y) \sum_{x \neq y} Q_t(y, x)\, D\big(\hat{s}(x, y, t), s(x, y, t)\big) + c_t,$$

where $c_t \in \mathbb{R}$ does not depend on s. Since $\hat{\mu}_t > 0$ on $\mathcal{X}$ (Lemma 7) and $D(a, b) = 0$ iff $a = b$, $\hat{L}_t^M$ is minimized, uniquely on the pairs with $Q_t(y, x) > 0$, at $s = \hat{s}$; likewise $\int_0^T \hat{L}_t^M(s)\mathrm{d}t$ is uniquely minimized at $s = \hat{s}$.

Proof. By (8) and (7), the s-dependent part of $\hat{L}_t(s)$ is $\sum_y \mu_t^{(n)}(y) \sum_{x \neq y} Q_t(y,x)[s - \frac{\mu_t^{(n)}(x)}{\mu_t^{(n)}(y)} \log s] = \sum_y \sum_{x \neq y} Q_t(y,x)[\mu_t^{(n)}(y)s - \mu_t^{(n)}(x) \log s]$ (a term with $\mu_t^{(n)}(y) = 0$ is 0 on both sides, as then $\mu_t^{(n)}(x) = 0$ for x reachable from y, cf. the proof of Lemma 12), and similarly for L_t^U with μ_t^U; the term $\phi(s_U^\star)$ in (7) does not depend on s. Both expressions are linear in the measure, and $\frac{n}{n+\gamma}\mu_t^{(n)} + \frac{\gamma}{n+\gamma}\mu_t^U = \hat{\mu}_t$ by linearity of $P_{0,t}$. Hence the s-dependent part of $\hat{L}_t^M(s)$ is $\sum_y \sum_{x \neq y} Q_t(y,x)[\hat{\mu}_t(y)s - \hat{\mu}_t(x) \log s] = \sum_y \hat{\mu}_t(y) \sum_{x \neq y} Q_t(y,x)[s - \hat{s} \log s]$, and $s - \hat{s} \log s = D(\hat{s}, s) + \hat{s} - \hat{s} \log \hat{s}$, where the last two terms do not depend on s. □

Proof of Lemma 5. The matrix Q_0 is in this case symmetric, so $P_{0,t}$ is doubly stochastic, in the sense that both columns and rows sum to 1. Furthermore, since the change of one token b_0 to b has probability $\frac{1-\alpha_t}{K} + \alpha_t \mathbf{1}_{b=b_0} \leqslant \frac{1}{K} + \alpha_t$, it follows that $P_{0,t}(x, x_0) \leqslant \frac{1}{K^d}(1 + K\alpha_t)^d$ for all $x, x_0 \in \mathcal{X}_0$. Using now this inequality, we have that $\mu_T(x) = \sum_{x_0} P_{0,T}(x, x_0)\mu_{\text{data}}(x_0) \leqslant \frac{1}{K^d}(1+K\alpha_T)^d$, so

$$\mathrm{KL}(\mu_T \mid \mathrm{Unif}(\mathcal{X}_0)) = \sum_x \mu_T(x) \log(K^d \mu_T(x)) \leqslant d \log(1 + K\alpha_T) \leqslant dK\alpha_T \leqslant \frac{1}{n}. \tag{66}$$

Notice that above we used that $\alpha_T = e^{-\bar{\sigma}_T} \leqslant (dKn)^{-1}$.

By Lemma 10, $M_\gamma^{-1} \leq \hat{s} \leq M_\gamma$ on $\mathcal{X}_0^2 \times [0,T]$ with $M = K^d(n+\gamma)/\gamma$, and $\hat{s}$ is continuous in t. Hence $|\hat{Q}_u(x,y)| \leq M_\gamma \sigma_{T-u} \max|Q_0|$, which is integrable on $[0,T]$ because $\bar{\sigma}_T < \infty$; so we have a unique absolutely continuous solution $(\nu_u)_{u \in [0,T]}$ on the closed interval, $\hat{\nu} = \nu_T$, and $\hat{\nu}_{T-T_0} \to \nu_T$ as $T_0 \downarrow 0$ by continuity. For every $T_0 \in (0,T)$ the hypotheses of Lemma 12 hold with $\mu_\infty = \mathrm{Unif}(\mathcal{X}_0) > 0$ and $s = \hat{s}$, and Lemma 11 gives $\int_{T_0}^T L_t(\hat{s})\,\mathrm{d}t \leq \mathrm{KL}(\mu_{\text{data}}|\hat{\mu}_0^{(\gamma)})$. Together with the bound $\mathrm{KL}(\mu_T|\mathrm{Unif}(\mathcal{X}_0)) \leq 1/n$ above, Lemma 12 yields

$$\mathrm{KL}(\mu_{T_0} \,|\, \nu_{T-T_0}) \leq \mathrm{KL}(\mu_{\text{data}} \,|\, \hat{\mu}_0^{(\gamma)}) + \frac{1}{n}. \tag{67}$$

Finally, $\mu_{T_0} = e^{\bar{\sigma}_{T_0} Q_0} \mu_{\text{data}} \to \mu_{\text{data}}$ as $T_0 \downarrow 0$ since $\bar{\sigma}_{T_0} \to 0$, and on a finite set $(\mu, \nu) \mapsto \mathrm{KL}(\mu|\nu)$ is jointly lower semicontinuous, being a supremum of continuous functions by the Donsker–Varadhan formula $\mathrm{KL}(\mu|\nu) = \sup_f\{\sum_x f(x)\mu(x) - \log \sum_x e^{f(x)}\nu(x)\}$. Hence

$$\mathrm{KL}(\mu_{\text{data}} \,|\, \hat{\nu}) \leq \liminf_{T_0 \downarrow 0} \mathrm{KL}(\mu_{T_0} \,|\, \nu_{T-T_0}) \leq \mathrm{KL}(\mu_{\text{data}} \,|\, \hat{\mu}_0^{(\gamma)}) + \frac{1}{n}.$$

□

Proof of Theorem 2. Fix a sample S; $\gamma(S) \in [1, \min\{n, K^d\}]$, $\hat{\mu}_0^{(\gamma(S))} > 0$ on $\mathcal{X}_0$ and $\hat{s}$ is its score, so Lemma 5 gives $\mathrm{KL}(\mu_{\text{data}}|\hat{\nu}) \leq \mathrm{KL}(\mu_{\text{data}}|\hat{\mu}_0^{(\gamma(S))}) + 1/n$. Taking expectations and using Lemma 2 gives item 1 of Theorem 1 with C replaced by $C+1$, because $1/n \leq k_{\text{eff}}/n$ ($k_{\text{eff}} \geq k_{\text{eff}}(\mu_{\text{data}}, 1) = 1$ for $n \geq 1$). Intersecting with the event A of Lemma 3, which does not depend on γ, gives item 2 in the same way. Corollaries 1,2,4 then transfer verbatim, as recorded in their proofs. □

D Experiment details

All experiments use the masked kernel with $\alpha_t = e^{-t}$ and $\kappa_t = \alpha_t/(1-\alpha_t)$. Seeds are fixed throughout.

Targets for label smoothing. We take $K = 4$ and $d = 8$, so $K^d = 65{,}536$. A $0/1$ matrix $A \in \{0,1\}^{K \times K}$ specifies the allowed token transitions. The first token is drawn from the stationary law of the chain, and each later token from the row-normalized weights, which are $\mathrm{Dirichlet}(2)$ on the allowed entries. The support is therefore exactly the set of length-d paths through A, and $l = \mathbf{1}^\top A^{d-1}\mathbf{1}$. We enumerate all 2^{16} matrices and keep the $25{,}696$ that are strongly connected and give every state at least one successor. For each target $l \in \{16, 64, 256, 1024, 4096\}$ and each seed, we draw A uniformly among those with l within $\pm 5\%$ of the target. We use $n \in \{64, 256, 1024\}$ and 5 seeds, each seed fixes A, the weights and the sample, and the same sample is used for every γ.

Label-smoothed estimator. By Lemma 4, the masked sampler outputs exactly $\hat{\mu}_0^{(\gamma)}$. We therefore compute $\mathrm{KL}(\mu_{\mathrm{data}} \mid \hat{\mu}_0^{(\gamma)})$ exactly by summing over $\mathcal{V}^d$, with no simulation.

Clipped empirical score. For the masked kernel, the estimator of Zhang et al. (2026, Eq. (11)) reads $\hat{s}_t(x, y) = \mathrm{clip}(\kappa_t \psi_c^{(n)}(a), \varepsilon, \kappa_t)$, and we set $\hat{s}_t = \varepsilon$ on unseen contexts ($n_c = 0$). Following their Theorem 3.8, we use early stopping at $t_0 = 1/n$, terminal time $T = \log n$, and $\varepsilon = c_0 \log n/n$, with $c_0 \in \{0.1, 1, 10\}$ chosen per (l, n) using the true KL.

Since the floor ε/κ_t depends on t, this score is improper, and the random-order sampler is not exact for it. We instead compute its output law by integrating the reverse Kolmogorov equation over all $(K + 1)^d = 390{,}625$ partially masked states. We use RK4 on 400 log-spaced times from T to t_0. Positions still masked at t_0 are unmasked one at a time in uniformly random order, with each token drawn with probability $\propto \hat{s}_{t_0}(\cdot, y)$. As a check, running the same solver with the proper score $\kappa_t \hat{\psi}_c$ reproduces $\hat{\mu}_0^{(\gamma)}$ with a maximum relative error of 7.2×10^{-6}.

Training. We use the sparse Markov chain of Zhang et al. (2026, App. E): $K = 100$, $d = 12$, 20 significant successors per state, $\epsilon_{\mathrm{bg}} = 0.01$ and $p_{\min} = 10^{-3}$. Held-out sequences are drawn fresh from the same chain. All training sequences are distinct, so $D_n = n$.

The model is their transformer (4 layers, width 128, 4 heads, ≈ 0.83M parameters). We train it with AdamW (learning rate 5×10^{-4}, $\beta = (0.9, 0.999)$, weight decay 10^{-3}, batch size 128). To train on $\hat{\mu}_0^{(\gamma)}$, each sequence is replaced by a uniform element of $\mathcal{V}^d$ with probability $p = \gamma/(n + \gamma)$, for $\gamma \in \{0, 64, 256, 1024, 4096\}$ and $n \in \{1024, 2896, 8192\}$. Each arm runs for $\lceil 6000/(1 - p) \rceil$ iterations, so that all arms see the same expected number of real sequences. We evaluate the held-out NELBO at 40 log-spaced iterations starting at 10. Each of the 5 seeds fixes both the sample and the initialization.