Error Analysis of Masked Discrete Diffusion under Low Complexity Structure

Anonymous authors
Paper under double-blind review

Abstract

Masked Diffusion Models have emerged as a discrete analogue of continuous Gaussian diffusion. They provide a new framework for learning distributions on discrete spaces and an alternative to autoregressive models for text generation. Despite the strong empirical performance of these models, theoretical analysis has been limited, focusing primarily on score estimation error, sampling guarantees and studying the structure of its Markov kernel. In this paper, we take a step towards end-to-end error analysis, derive approximation and finite-sample statistical error bounds for a Transformer-based predictor, analogous to the corresponding theory of continuous diffusion models. Our analysis identifies structural conditions on the clean-data conditional distributions under which their logits can be efficiently approximated by Transformers, and combines this approximation guarantee with Transformer complexity bounds to control estimation error. We further test empirically whether these structural conditions are supported by some common discrete distributions. This provides the first approximation-estimation theory for Transformer-based Masked Diffusion Models trained with a continuous-time NELBO loss and a data-dependent scaling law for the Transformer's attention window.

1 Introduction

Classical diffusion models have achieved great success in image generation, audio synthesis, time-series modelling and other settings where data admits a continuous representation. Recently, masked diffusion models have emerged as a discrete analogue to continuous diffusion (Austin et al., 2021; Sahoo et al., 2024; Shi et al., 2024; Ou et al., 2025). These models define a forward process that gradually replaces tokens in a sequence with a special mask token, until the sample becomes fully masked. A neural network is then trained to predict the clean token sequence from a partially masked sequence, effectively learning the posterior distribution of the clean token sequence given the currently visible context. At generation time, sampling begins from the all-mask sequence and repeatedly applies the learned reverse process to replace masked tokens with predicted tokens, eventually producing a clean sample.

Masked diffusion models have recently been adopted in practice for language modelling (He et al., 2023; Zheng et al., 2024; Lou et al., 2024; Nie et al., 2025; Zhao et al., 2025; Ye et al., 2025; Yang et al., 2025). On the theoretical side, some works study the structure of the continuous-time generator and discrete score (Campbell et al., 2022; Benton et al., 2024; Ou et al., 2025). Recent studies have also heavily focused on sampling guarantees (Li & Cai, 2025; Chen & Ying, 2025; Zhang et al., 2025; Chen et al., 2026; Liang et al., 2025a; Huang et al., 2025; Ren et al., 2025a;b; Liang et al., 2025b; Pham et al., 2025). On the statistical learning theory side, Wakasugi & Suzuki (2025); Srikanth et al. (2026); Zhang et al. (2026) address score-estimation error. These developments motivate a closer study of the connection between the structure of the conditional distributions to be learned and the expressive power and complexity of the neural predictor. In particular, how should the predictor's size and context length depend on this structure and on the number of training samples?

In this paper, we address this question using Transformers as predictors of the tokenwise conditional distributions. To make the role of context length explicit, we consider a one-block sliding-window architecture, in which each prediction uses only a local window around the masked position. We identify structural conditions under which such predictors can efficiently approximate the full-context conditional distributions. Specifically, we assume that the smoothed conditional logits admit

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.

representations with bounded interaction order and exponentially decaying contributions from distant coordinates. These conditions control both the error from restricting the context and the complexity of representing the remaining within-window interactions.

Combining the resulting approximation bounds with Transformer complexity bounds, we obtain finite-sample bounds on the excess continuous-time NELBO risk of an empirical-risk minimizer. These bounds also control the KL divergence between the data distribution and the distribution generated by the learned continuous-time reverse process, with the smoothing bias kept explicit. The resulting bounds quantify how the context window and model size can be chosen to balance approximation and estimation errors, depending on the structure of the conditional distributions and the sample size. Finally, we examine the spatial decay and interaction-order properties empirically for Ising and General Time Reversible (GTR) models.

2 Related Work

Wakasugi & Suzuki (2025) derive finite-sample bounds for estimating the discrete score $s_t(x, y) = \frac{p_t(y)}{p_t(x)}$. Their result is related in spirit to ours, since both analyses exploit structure of the data distribution to avoid a worst-case dependence on the size of the discrete state space. However, their assumptions do not cover the absorbing process used in masked diffusion. In particular, they assume that the transition-rate matrix is symmetric. Masked diffusion instead contains one-way transitions from ordinary tokens to the mask state and hence does not satisfy this condition.

Zhang et al. (2026) analyze the finite-sample statistical error for score estimation and compare how different transition kernels adapt to low complexity structure on the data. However, they do not impose a parametric class on the neural network, and hence do not address how approximation error can be controlled in the presence of such low complexity structure.

Srikanth et al. (2026) attempt an end-to-end analysis of discrete diffusion models, including both the *approximation error* and the *finite-sample statistical error*. The proof of their Lemma 1 implies that approximation error can be made 0 once the network has enough parameters to represent an arbitrary function defined on $\mathcal{S}^d$, without exploiting any structure of the clean data distribution. In contrast, our analysis assumes that the clean data distribution satisfies an explicit structural condition restricting its complexity, and derives an approximation bound whose size depends on that complexity.

Wan et al. (2026) and Su et al. (2025) derive approximation and finite sample statistical error bounds in the setting of discrete flow. Their methods can technically be used to give bounds in the setting of masked discrete diffusion. However, we do not follow this route primarily because they assume that the reverse rate is extendible to a Hölder continuous function on the Euclidean space. While such an assumption is trivially satisfied for discrete functions, it is difficult to interpret and measure the parameters of the Hölder class in terms of the properties of the distribution. Instead, our analysis uses conditions which have a simple probabilistic interpretation for common data distributions. Secondly, they use a squared l_2 loss, while we perform our analysis for the continuous-time NELBO loss which is more commonly used in Discrete Diffusion.

3 Approximation theory of sliding window Transformer

Model. We use a one-block sliding-window Transformer as a predictor; each token is first embedded and augmented with a positional encoding PE, then passed through multi-head attention and a position-wise feed-forward network as usual. The only nonstandard feature is that attention at position j is restricted to the local window $\{j-U, \dots, j+U\}$, where U is the half-window size. Thus, if $Z = (z_j)_{j\in\mathbb{Z}}$ denotes the encoded sequence and $Z_{j,U} := [z_{j-U}, \dots, z_{j+U}]$, then

$$\mathsf{Attn}_{U,H}(Z)_j = z_j + \sum_{h=1}^{H} W_h^V Z_{j,U} \,\mathrm{Softmax}\Big((W_h^K Z_{j,U})^\top W_h^Q z_j\Big),$$

so the output at position j depends only on tokens within distance U of j. The final output is obtained by applying the same position-wise feed-forward network g: $\widehat{F}_j(X) = (\mathsf{FNN}_g \circ \mathsf{Attn}_{U,H} \circ \mathrm{Enc}_{\mathrm{PE}})(X)_j$. See Appendix A.2 for details. For the results below, we use D for the embedding dimension and H for the number of heads; the position-wise FNN has depth L_{net}, maximal hidden

width W_{net}, and sparsity S_{net} (the total number of nonzero affine parameters). We write B_{net} for the largest absolute scalar network parameter and $B_{\mathrm{cov}} := \max\{1, B_{\mathrm{net}}, \|\mathrm{PE}\|_\infty\}$. For this architecture, we will later use the covering-number estimate of Takakura & Suzuki (2023). See Appendix A.3.

Function class and regularity. To develop an approximation theory for this Transformer, we first fix a class of functions, inspired by Takakura & Suzuki (2023), but tailored to discrete spaces.

Let $E = \{0, \dots, q-1\}$ carry a full-support probability measure ν, and let $\mu = \nu^{\otimes \mathbb{Z}}$ on $\mathsf{X}_E = E^{\mathbb{Z}}$. Choose an orthonormal basis $\{\varphi_0, \dots, \varphi_{q-1}\}$ of $L^2(E, \nu)$ with $\varphi_0 \equiv 1$. For each finitely supported $\gamma \in \mathfrak{A} \subset \{0, \dots, q-1\}^{\mathbb{Z}}$ and $X \in \mathsf{X}_E$, set

$$\chi_\gamma(X) := \prod_{j \in \operatorname{supp} \gamma} \varphi_{\gamma_j}(X_j), \qquad b_\gamma := \frac{\chi_\gamma}{\|\chi_\gamma\|_\infty}, \qquad d_\gamma := \widehat{f}(\gamma)\|\chi_\gamma\|_\infty.$$

Then $\{\chi_\gamma\}_{\gamma \in \mathfrak{A}}$ is an orthonormal basis of $L^2(\mathsf{X}_E, \mu)$ and $f = \sum_{\gamma \in \mathfrak{A}} d_\gamma b_\gamma$, with $\|b_\gamma\|_\infty \le 1$.

For $\gamma \in \mathfrak{A}$, define

$$\operatorname{ord}(\gamma) := |\operatorname{supp} \gamma|, \qquad \varrho(\gamma) := \max\{|j| : \gamma_j \neq 0\},$$

with $\varrho(0) := 0$. For $\beta > 0$, $1 \le \tau \le \infty$, and $r \in \mathbb{N}$, define

$$\mathcal{B}^{(r)}_{\beta,\tau} := \left\{ f = \sum_\gamma d_\gamma b_\gamma : \left\| \left(2^{\beta \varrho(\gamma)} |d_\gamma|\right)_\gamma \right\|_{\ell^\tau} < \infty,\ d_\gamma = 0 \text{ if } \operatorname{ord}(\gamma) > r \right\}.$$

Intuitively, $\mathcal{B}^{(r)}_{\beta,\tau}$ consists of functions built from interaction terms, each involving at most r coordinates, with exponentially small coefficients with decay rate β. We denote the displayed ℓ^τ quantity by $\|f\|_{\mathcal{B}^{(r)}_{\beta,\tau}}$. For $\tau = 1$, this coefficient norm also has a familiar metric interpretation: finiteness of this norm implies Lipschitzness in the shift-space metric; see Appendix A.4. Let $(\Sigma_j X)_k := X_{k+j}$ and $F^\circ_j := f \circ \Sigma_j$. Since μ is shift-invariant, $F^\circ = (F^\circ_j)_{j \in \mathbb{Z}}$ is shift-equivariant.

4 Theoretical results - sliding window transformer

Let τ' be the Hölder conjugate of τ ($\tau' = \infty$ for $\tau = 1$ and $\tau' = 1$ for $\tau = \infty$), and set $\kappa := \frac{r-1}{\tau'}$.

Theorem 4.1 (Transformer approximation error)**.** *Let $f \in \mathcal{B}^{(r)}_{\beta,\tau}$ satisfy $\|f\|_{\mathcal{B}^{(r)}_{\beta,\tau}} \le 1$ and $F^\circ_j = f \circ \Sigma_j$. For every $K \ge 2$, there exists a one-block sliding-window Transformer $\widehat{F}_K$ such that*

$$\sup_{j \in \mathbb{Z}} \|\widehat{F}_{K,j} - F^\circ_j\|_\infty \le C(K+1)^\kappa 2^{-\beta K}.$$

It can be chosen with $U_K = K$, $H_K = 2K+1$, $D_K = q + q(2K+1) + 2$, and

$$L_{\mathrm{net},K} \lesssim K+1, \quad W_{\mathrm{net},K} \lesssim (K+1)^r, \quad S_{\mathrm{net},K} \lesssim (K+1)^{r+1}, \quad B_{\mathrm{net},K} \lesssim (K+1)^3 \log(K+2),$$

with number of parameters $N_K \lesssim (K+1)^{r+1} + q^2(K+1)^3$. The implicit constants depend only on q, ν, r, β, τ.

Two direct consequences translate this approximation bound into network size bounds; we state them in Appendix A.5. Proof is deferred to Appendix B.

5 Masked Diffusion Models

Fix an alphabet $\mathcal{A} = [M]$ ($M \ge 2$), sequence length $L \ge 1$, data law P on $\mathcal{A}^L$. With mask symbol $\mathtt{m} \notin \mathcal{A}$, enlarge the alphabet to $\mathcal{E} := \mathcal{A} \cup \{\mathtt{m}\}$, $|\mathcal{E}| = M+1$. For $z \in \mathcal{E}^L$ write $\mathcal{M}(z) := \{i : z_i = \mathtt{m}\}$ and $\mathcal{O}(z) := [L] \setminus \mathcal{M}(z)$ for its set of masked and observed coordinates respectively.

Choose a strictly decreasing C^1 function $\alpha : [0,1] \to [0,1]$ with $\alpha_0 = 1$ and $\alpha_1 = 0$. Following Austin et al. (2021); Campbell et al. (2022), we formulate masked diffusion as a time inhomogeneous Continuous Time Markov Chain (CTMC) $Z = (Z_t)_{t \in [0,1]}$ on $\mathcal{E}^L$, initialized at $Z_0 = X \sim P$.

Conditioned on $X = x$, the coordinates of Z_t evolve independently. At time t, each token is independently retained with probability α_t and masked otherwise. Hence $Z_1 = \mathtt{m}^L$ almost surely and

$$q_t(z \mid x) := \mathbb{P}(Z_t = z \mid X = x) = \prod_{j=1}^{L} \Big(\alpha_t \mathbf{1}_{\{z_j = x_j\}} + (1 - \alpha_t) \mathbf{1}_{\{z_j = \mathtt{m}\}} \Big), \tag{5.1}$$

$$q_t(z) = \sum_{x \in \mathcal{A}^L} P(x) q_t(z \mid x), \qquad w(t) := -\frac{\alpha_t'}{1 - \alpha_t}, \quad t \in (0, 1]. \tag{5.2}$$

A masking-time construction of the CTMC is given in Appendix A.1.

The true reverse process starts from $\mathtt{m}^L$ at $t = 1$ and unmasks coordinates as t decreases. For $i \in \mathcal{M}(z)$ and $a \in \mathcal{A}$, write $z^{i \leftarrow a}$ for z with coordinate i replaced by a. The derivation of the rate of the reverse process is given in Appendix C.2; here we record the quantities needed below.

Denote by $\mathcal{Z}_P := \{z \in \mathcal{E}^L : \exists x \in \mathcal{A}^L,\ P(x) > 0,\ z_j \in \{x_j, \mathtt{m}\}\ \forall j\}$ the set of states reachable by the forward process with positive probability. Since the masking times are independent of X, we have that $X \mid Z_t = z$ has the same law as $X \mid X_{\mathcal{O}(z)} = z_{\mathcal{O}(z)}$ for all t and reachable z. Thus, for $z \in \mathcal{Z}_P$ and $i \in \mathcal{M}(z)$, define the posterior

$$\pi_i(a \mid z) := \mathbb{P}\big(X_i = a \mid X_{\mathcal{O}(z)} = z_{\mathcal{O}(z)}\big). \tag{5.3}$$

The rate of the true reverse process is $r_t(z, z^{i \leftarrow a}) = w(t) \pi_i(a \mid z)$, while replacing the posterior by a neural predictor $p_i(\cdot \mid z, t)$ gives the learned rate $r_t^p(z, z^{i \leftarrow a}) = w(t) p_i(a \mid z, t)$. Thus $w(t)$ controls *when* a coordinate is unmasked, whereas $\pi_i(\cdot \mid z)$ controls *which token* is revealed.

5.1 Problem Formulation

Given i.i.d. clean sequences from P, the predictor is trained with the continuous-time NELBO loss, commonly used in masked diffusion models (Sahoo et al., 2024):

$$R(p) := \mathbb{E}_{X \sim P} \int_0^1 w(t) \sum_{z \in \mathcal{E}^L} q_t(z \mid X) \sum_{i \in \mathcal{M}(z)} -\log p_i(X_i \mid z, t)\, dt. \tag{5.4}$$

We introduce some notation for path measures. First, let $\Omega := D([0, 1], \mathcal{E}^L)$ be the space of paths and $\Phi(\omega) := \omega(0)$ for every path ω. Denote the true reverse-path law by $\mathbb{J}_\Omega$ ($\Phi_\# \mathbb{J}_\Omega = P$), the learned law by $\mathbb{P}_p^{(\infty)}$. Denote its endpoint law as $P_p^{(\infty)} := \Phi_\# \mathbb{P}_p^{(\infty)}$. In practice, we consider a sampler running for finitely many steps. Hence, we define T-step laws by $\mathbb{P}_p^{(T)}$ and $P_p^{(T)} := \Phi_\# \mathbb{P}_p^{(T)}$. The error from finite-step sampling is

$$\mathrm{Disc}_p(T) := \Big[D_{\mathrm{KL}}(\mathbb{J}_\Omega \| \mathbb{P}_p^{(T)}) - D_{\mathrm{KL}}(\mathbb{J}_\Omega \| \mathbb{P}_p^{(\infty)}) \Big]_+ .$$

Our goal is to control the errors $D_{\mathrm{KL}}(P \| P_p^{(\infty)})$ and $D_{\mathrm{KL}}(P \| P_p^{(T)})$.

5.2 Assumptions

The posterior $\pi_i(a \mid z)$ can vanish, whereas our approximation theory uses bounded real-valued logits. To deal with this, fix $\lambda \in (0, 1)$ and smooth the posterior:

$$\pi_i^\lambda(a \mid z) := (1 - \lambda) \pi_i(a \mid z) + \frac{\lambda}{M}, \qquad \ell_{i,a}^\lambda(z) := \log \pi_i^\lambda(a \mid z), \tag{5.5}$$

$$A_\lambda := \log(M/\lambda), \qquad -A_\lambda \le \ell_{i,a}^\lambda(z) \le 0. \tag{5.6}$$

The cost of this smoothing will remain explicit in our final bound.

5.2.1 Regularity and shift equivariance

The posterior logits are functions on the finite space $\mathcal{E}^L$, whose domain depends on the sequence length L. To obtain the L-independent bounds below, we represent them as functions on a common

space of bi-infinite sequences. Introduce a padding symbol $\partial \notin \mathcal{E}$, and set $\widetilde{\mathcal{E}} := \mathcal{E} \cup \{\partial\}$ and $Q := |\widetilde{\mathcal{E}}| = M + 2$. For $z \in \mathcal{E}^L$, let $\iota z \in \widetilde{\mathcal{E}}^{\mathbb{Z}}$ agree with z on $[L]$ and equal ∂ outside $[L]$. We write

$$(\Sigma_i Y)_k := Y_{k+i}, \qquad Y \in \widetilde{\mathcal{E}}^{\mathbb{Z}},$$

so that $\Sigma_i \iota z$ moves position i of the finite sequence to a common reference position.

We can now apply the regularity framework of Section 3 on the common space $\widetilde{\mathcal{E}}^{\mathbb{Z}}$. Choose a full-support probability measure ν on $\widetilde{\mathcal{E}}$, and specialize the notation there to $E = \widetilde{\mathcal{E}}$, $q = Q$, and reference measure ν. In particular, we use the class $\mathcal{B}^{(r)}_{\beta,\tau}$ in Section 3 to impose spatial decay and bounded interaction order on the posterior logits. We make this precise in the following assumption.

Assumption 5.1. For every $a \in \mathcal{A}$ there is $f^\lambda_a \in \mathcal{B}^{(r)}_{\beta,\tau}$ such that

$$\max_{a \in \mathcal{A}} \|f^\lambda_a\|_{\mathcal{B}^{(r)}_{\beta,\tau}} \le R_\lambda, \tag{5.7}$$

and, for every $z \in \mathcal{Z}_P$, $i \in \mathcal{M}(z)$ and $a \in \mathcal{A}$,

$$\ell^\lambda_{i,a}(z) = f^\lambda_a(\Sigma_i \iota z). \tag{5.8}$$

Here R_λ may depend on the data through $P, \lambda, M, \nu, \beta, \tau, r$, but not on L. The exponential weight encodes rapid decay of distant interactions and $\mathrm{ord}(\gamma) \le r$ bounded interaction order, as motivated by high-temperature graphical models; Section 7 tests both conditions for Ising and GTR models. Appendix G treats polynomial spatial decay, relevant when dependence is longer-ranged. Shift equivariance is motivated by stationarity and permits the same feed-forward network at every position.

5.3 One-block Sliding-Window Transformer

We now specialize the one-block Transformer of Section 3 to the padded inputs $\iota z \in \widetilde{\mathcal{E}}^{\mathbb{Z}}$. For M scalar branches $\mathbf{F} = (F_a)_{a \in \mathcal{A}}$, their softmax defines

$$p_{\mathbf{F},i}(a \mid z, t) = p_{\mathbf{F},i}(a \mid z) := \frac{e^{F_{a,i}(\iota z)}}{\sum_{b \in \mathcal{A}} e^{F_{b,i}(\iota z)}}. \tag{5.9}$$

Let $\mathcal{T}_K := \mathcal{T}(1, U_K, D_K, H_K, L_{\mathrm{net},K}, W_{\mathrm{net},K}, S_{\mathrm{net},K}, B_{\mathrm{cov},K}; \mathrm{PE}_K)$ be the corresponding K-dependent scalar class with fixed positional encoding and the uniform norm of Takakura & Suzuki (2023, Theorem 5.3). Restriction to padded inputs and coordinate projection do not enlarge this norm; see Appendix A.6.

With $\rho_K := C_{\mathrm{tr}} R_\lambda (K+1)^\kappa 2^{-\beta K}$, define

$$\mathcal{P}^{\mathrm{bd}}_{K,\lambda} := \left\{ p_{\mathbf{F}} : \mathbf{F} \in \mathcal{T}_K^M,\ -A_\lambda - \rho_K \le F_{a,i}(\iota z) \le \rho_K \ \ \forall\, z \in \mathcal{Z}_P,\ i \in \mathcal{M}(z),\ a \in \mathcal{A} \right\}. \tag{5.10}$$

For p in this class, let $J_p(x)$ be the conditional loss obtained by fixing $X = x$, so that $R(p) = \mathbb{E} J_p(X)$; the explicit formula is recorded in Appendix A.7. Finally, define the loss envelope $\mathcal{B}_{K,\lambda} := L(A_\lambda + 2\rho_K + \log M)$ and the complexity quantities $V_K := C_{\mathrm{TS}} M L_{\mathrm{net},K}(S_{\mathrm{net},K} + H_K D_K^2)$ and $C_{0,K} := e D_K H_K L_{\mathrm{net},K} W_{\mathrm{net},K} B_{\mathrm{cov},K}$.

6 Theoretical results - Masked Diffusion Models

We first connect the learning objective to the quantity of interest, namely the KL divergence between the data distribution and the generated distribution.

Theorem 6.1 (Population risk controls generation error)**.** *Assume $R(p) < \infty$ and that $p_i(a \mid z, t) > 0$ for all $i \in [L]$, $a \in \mathcal{A}$. Then*

$$D_{\mathrm{KL}}\Big(P \Big\| P_p^{(\infty)}\Big) \le R(p) - H(P).$$

Moreover, if $D_{\mathrm{KL}}(\mathbb{J}_\Omega \| \mathbb{P}_p^{(T)}) < \infty$, then

$$D_{\mathrm{KL}}\Big(P \Big\| P_p^{(T)}\Big) \le R(p) - H(P) + \mathrm{Disc}_p(T).$$

Theorem 6.1 reduces the generation problem to controlling excess population risk. For a learned predictor, this excess risk admits the usual decomposition into approximation and estimation errors. For a predictor class $\mathcal{P}_K$ and a data-dependent estimator $\widehat{p}_{K,m} \in \mathcal{P}_K$, define

$$\mathrm{Approx}(\mathcal{P}_K) := \inf_{p \in \mathcal{P}_K} \big(R(p) - H(P)\big), \qquad \mathrm{Est}_{K,m} := R(\widehat{p}_{K,m}) - \inf_{p \in \mathcal{P}_K} R(p).$$

Thus Theorem 6.1 turns generation error into the sum of approximation and estimation errors (plus $\mathrm{Disc}_{\widehat{p}_{K,m}}(T)$ for a finite-step sampler); the formal statement is Corollary A.4 in the appendix.

We next control approximation. Restricting to $\mathcal{P}^{\mathrm{bd}}_{K,\lambda}$ introduces two errors: smoothing the posterior by λ to obtain bounded logits, and truncating dependence to a radius-K window. Define

$$\rho_{\lambda,\mathrm{Tr}}(K) := C_{\mathrm{tr}} R_\lambda (K+1)^\kappa 2^{-\beta K}, \qquad K \geq 2. \tag{6.1}$$

Theorem 6.2 (Posterior approximation error)**.** *Suppose that Assumption 5.1 holds. For every integer $K \geq 2$, the class $\mathcal{P}^{\mathrm{bd}}_{K,\lambda}$ contains a time-independent predictor p_K satisfying*

$$R(p_K) - H(P) \leq L \log \frac{1}{1-\lambda} + 2L\rho_{\lambda,\mathrm{Tr}}(K), \tag{6.2}$$

$$\mathrm{Approx}(\mathcal{P}^{\mathrm{bd}}_{K,\lambda}) \leq L \log \frac{1}{1-\lambda} + 2LC_{\mathrm{tr}} R_\lambda (K+1)^\kappa 2^{-\beta K}. \tag{6.3}$$

Every scalar branch may be chosen with

$$\begin{gathered} U_K = K, \quad H_K = 2K+1, \quad D_K = Q + Q(2K+1) + 2, \\ L_{\mathrm{net},K} \leq C_{\mathrm{tr}}(K+1), \quad W_{\mathrm{net},K} \leq C_{\mathrm{tr}}(K+1)^r, \quad S_{\mathrm{net},K} \leq C_{\mathrm{tr}}(K+1)^{r+1}, \\ B_{\mathrm{net},K} \leq C_{\mathrm{tr}}(1 \vee R_\lambda)(K+1)^3 \log(K+2), \\ N_{K,M} \leq C_{\mathrm{tr}} M\big((K+1)^{r+1} + Q^2(K+1)^3\big). \end{gathered} \tag{6.4}$$

Here $N_{K,M}$ is the parameter count of the M-branch block-parallel realization.

The two terms are respectively the smoothing cost and the finite-window approximation cost; the latter decays exponentially in K up to $(K+1)^\kappa$.

Empirical risk and exact ERM. The preceding theorem is population-level. For i.i.d. $X^{(1)}, \ldots, X^{(m)} \sim P$, define

$$\widehat{R}_m(p) = \frac{1}{m} \sum_{j=1}^{m} J_p(X^{(j)}), \qquad \widehat{p}_m \in \arg\min_{p \in \mathcal{P}^{\mathrm{bd}}_{K,\lambda}} \widehat{R}_m(p). \tag{6.5}$$

We assume this exact ERM is measurable and exists.

Main result.

Theorem 6.3 (Excess-risk bound)**.** *Assume the alphabet, data law, masking process, and risk of Section 5; $\lambda \in (0,1)$, $K \geq 2$, and Assumption 5.1; the scalar branch class is a restriction/coordinate projection of the fixed-positional-encoding Takakura & Suzuki (2023) class with $B_{\mathrm{cov},K} \geq 1$ and $\|\mathrm{PE}_K\|_\infty \leq B_{\mathrm{cov},K}$; and $X^{(1)}, \ldots, X^{(m)} \overset{\text{i.i.d.}}{\sim} P$, with the measurable exact ERM* (6.5) *existing. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\begin{aligned} R(\widehat{p}_m) - H(P) \leq{}& L \log \frac{1}{1-\lambda} + 2L\rho_K + C_{\mathrm{ch}} \frac{\mathcal{B}_{K,\lambda}}{2} \sqrt{\frac{V_K}{m}} \\ &\times \left\{ \sqrt{\log\left[C_{0,K}\left(1 + \frac{8L}{\mathcal{B}_{K,\lambda}}\right)\right]} + \frac{\sqrt{\pi}}{2} \right\} + \mathcal{B}_{K,\lambda} \sqrt{\frac{2}{m} \log \frac{2}{\delta}}. \end{aligned} \tag{6.6}$$

Here $\rho_K, \mathcal{B}_{K,\lambda}, V_K, C_{0,K}$ are the quantities defined above, with H_K, D_K, Q as in Theorem 6.2 and Section 5.

Substituting the polynomial architecture bounds into Theorem 6.3 gives the simplified finite-sample inequality in Corollary A.5 of Appendix A.8, which we use only to expose the K-dependence below.

The bound also gives an explicit approximation–estimation tradeoff in the window size: under uniformly bounded structural parameters, balancing the two K-dependent terms gives $K_m = \frac{\log m}{2\beta \log 2} + O(\log\log m)$, independently of L; see Proposition A.6 in the appendix. The corresponding network has $N_{K_m,M} = O\big(M\big((\log m)^{r+1} + (M+2)^2 (\log m)^3\big)\big)$ parameters.

7 Experimental Setup

7.1 Ising Model

We consider the translation-invariant Ising model on $X = (X_1, \dots, X_L) \in \{-1, +1\}^L$,

$$P_T(x) = \frac{1}{Z_{T,L}} \exp\left\{ \frac{1}{T} \left(\sum_{i=1}^{L} \sum_{r=1}^{R_0} J_r x_i x_{i+r} + h \sum_{i=1}^{L} x_i \right) \right\}, \tag{7.1}$$

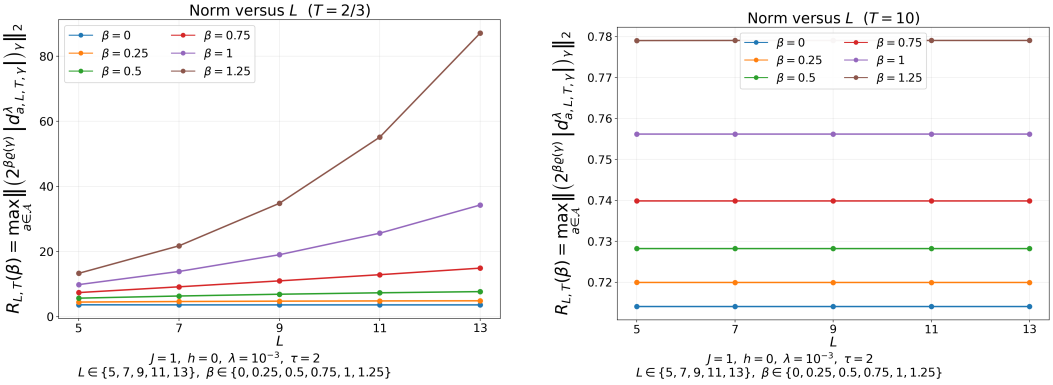
with periodic boundary conditions $x_{i+L} = x_i$. Here R_0 is the interaction range, $J_1, \dots, J_{R_0}$ are the interaction strengths, h is the external field, and $T > 0$ is the temperature.

7.1.1 Testing the norm condition

We take the nearest-neighbor model $R_0 = 1$, $J_1 = 1$, $h = 0$, and odd $L \in \{5, 7, 9, 11, 13\}$. Let $i_0 := (L+1)/2$ be the central site. We mask i_0 and enumerate all masked contexts on the remaining $L-1$ coordinates. For each $a \in \mathcal{A}$, we compute the posterior $\pi_{i_0,L,T}(a \mid z)$ exactly and, with $\lambda = 10^{-3}$, form $f^{\lambda}_{a,L,T}(z) = \log\big((1-\lambda)\pi_{i_0,L,T}(a \mid z) + \frac{\lambda}{2}\big)$. Using the expansion of Section 3, let $d^{\lambda}_{a,L,T,\gamma}$ denote the corresponding normalized coefficients. With $\tau = 2$, define

$$R_{L,T}(\beta) := \max_{a \in \mathcal{A}} \left[\sum_{\gamma} 2^{2\beta\varrho(\gamma)} \left| d^{\lambda}_{a,L,T,\gamma} \right|^2 \right]^{1/2}. \tag{7.2}$$

The regularity assumption of Section 5.2.1 predicts that, for suitable $\beta > 0$, $R_{L,T}(\beta)$ remains bounded as L increases. Figure 1 shows $L \mapsto R_{L,T}(\beta)$ for $T = 2/3$ and $T = 10$.



(a) $T = 2/3$. (b) $T = 10$.

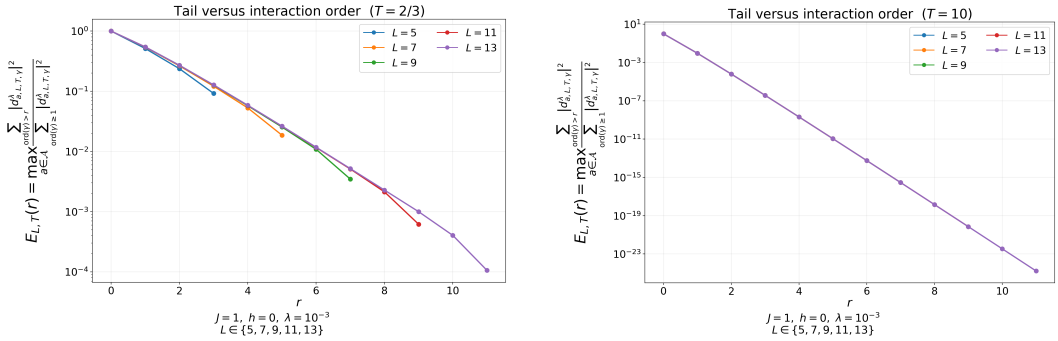
Figure 1: Testing assumptions for the Ising model: Norm as a function of sequence length.

7.1.2 Testing the interaction-order condition

Using the same coefficients, recall that $\mathrm{ord}(\gamma) = |\mathrm{supp}(\gamma)|$. For a cutoff r, define the fraction of nonconstant coefficient mass carried by interactions of order greater than r by

$$E_{L,T}(r) := \max_{a \in \mathcal{A}} \frac{\sum_{\mathrm{ord}(\gamma) > r} \left| d^{\lambda}_{a,L,T,\gamma} \right|^2}{\sum_{\mathrm{ord}(\gamma) \geq 1} \left| d^{\lambda}_{a,L,T,\gamma} \right|^2}. \tag{7.3}$$

The exact finite-interaction-order assumption would give $E_{L,T}(r) = 0$ beyond a fixed r; more generally, rapid decay of $E_{L,T}(r)$ indicates that high-order interactions contribute little. Figure 2 shows the tails $r \mapsto E_{L,T}(r)$ for $T = 2/3$ and $T = 10$ across the tested sequence lengths.



(a) $T = 2/3$. (b) $T = 10$.

Figure 2: Testing assumptions for the Ising model: Tail as a function of r.

7.2 General Time Reversible Model

We consider the four-state General Time Reversible (GTR) model on $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. Let $\pi = (\pi_{\mathrm{A}}, \pi_{\mathrm{C}}, \pi_{\mathrm{G}}, \pi_{\mathrm{T}})$ be a strictly positive stationary distribution and let $s_{ab} = s_{ba} > 0$ denote the exchangeability parameters. The generator G is defined by

$$G_{ab} = s_{ab}\pi_b, \qquad a \neq b, \qquad G_{aa} = -\sum_{b \neq a} s_{ab}\pi_b. \tag{7.4}$$

For $\eta > 0$, let $P_\eta := \exp(\eta G)$. Initializing the chain from π, the distribution of $x = (x_1, \ldots, x_L) \in \mathcal{A}^L$ is

$$P_{\eta,L}(x) = \pi_{x_1} \prod_{i=1}^{L-1} [P_\eta]_{x_i x_{i+1}}. \tag{7.5}$$

The symmetry of the exchangeability parameters implies detailed balance, $\pi_a[P_\eta]_{ab} = \pi_b[P_\eta]_{ba}$, so the chain is reversible with stationary distribution π.

For the experiments, we take $\pi = (0.3, 0.2, 0.2, 0.3)$, $s_{\mathrm{AG}} = s_{\mathrm{CT}} = 4$, and $s_{\mathrm{AC}} = s_{\mathrm{AT}} = s_{\mathrm{CG}} = s_{\mathrm{GT}} = 1$, which gives

$$G = \begin{pmatrix} -1.3 & 0.2 & 0.8 & 0.3 \\ 0.3 & -1.7 & 0.2 & 1.2 \\ 1.2 & 0.2 & -1.7 & 0.3 \\ 0.3 & 0.8 & 0.2 & -1.3 \end{pmatrix}.$$

7.2.1 Testing the norm condition

We use odd sequence lengths $L \in \{3, 5, 7, 9, 11\}$ and mask the central position. For each masked context z and $a \in \mathcal{A}$, we compute the exact posterior $\pi_{L,\eta}(a \mid z)$ and, with $\lambda = 10^{-3}$, form $f^\lambda_{a,L,\eta}(z) = \log\big((1-\lambda)\pi_{L,\eta}(a \mid z) + \frac{\lambda}{4}\big)$. Using the expansion of Section 3, let $d^\lambda_{a,L,\eta,\gamma}$ denote the corresponding normalized coefficients. With $\tau = 2$, define

$$R_{L,\eta}(\beta) := \max_{a \in \mathcal{A}} \left[\sum_{\gamma} 2^{2\beta\varrho(\gamma)} \left| d^\lambda_{a,L,\eta,\gamma} \right|^2 \right]^{1/2}. \tag{7.6}$$

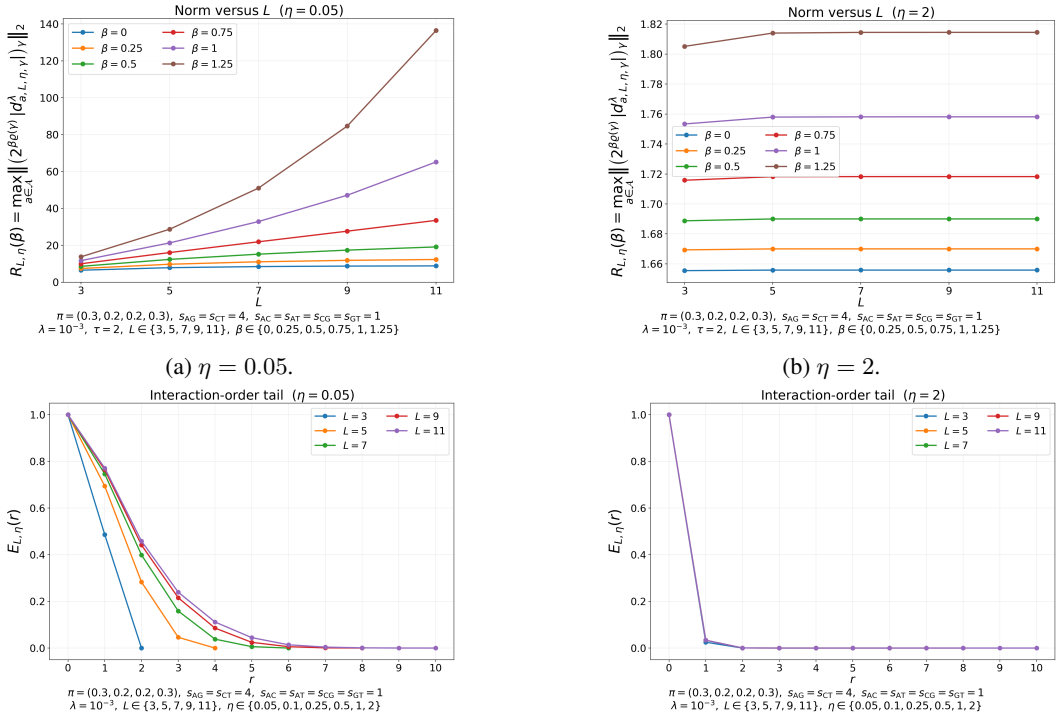
The regularity assumption predicts that, for suitable $\beta > 0$, $R_{L,\eta}(\beta)$ remains bounded as L increases. Figure 3 shows $L \mapsto R_{L,\eta}(\beta)$ for $\eta = 0.05$ and $\eta = 2$.

7.2.2 Testing the interaction-order condition

Using the same coefficients, recall that $\mathrm{ord}(\gamma) = |\operatorname{supp}(\gamma)|$. For a cutoff r, define

$$E_{L,\eta}(r) := \max_{a \in \mathcal{A}} \frac{\displaystyle\sum_{\mathrm{ord}(\gamma) > r} \left| d^\lambda_{a,L,\eta,\gamma} \right|^2}{\displaystyle\sum_{\mathrm{ord}(\gamma) \geq 1} \left| d^\lambda_{a,L,\eta,\gamma} \right|^2}. \tag{7.7}$$

The exact finite-interaction-order assumption would give $E_{L,\eta}(r) = 0$ beyond a fixed r; more generally, rapid decay of $E_{L,\eta}(r)$ indicates that higher-order interactions carry little coefficient mass. Figure 3 shows $r \mapsto E_{L,\eta}(r)$ for $\eta = 0.05$ and $\eta = 2$ across the tested sequence lengths.



(a) $\eta = 0.05$.

(b) $\eta = 2$.

(c) $\eta = 0.05$.

(d) $\eta = 2$.

Figure 3: Testing assumptions for the GTR model. Top: norm as a function of sequence length. Bottom: tail as a function of r.

8 Results and Discussion

Figure 1a shows that at temperature $T = 2/3$, $R_{L,T}(\beta)$ stabilizes with L for small values of the smoothness β (specifically, $0 \leq \beta \leq 0.25$ in the figure), whereas it blows up with L for larger values. On the other hand, Figure 1b shows that at temperature $T = 10$, $R_{L,T}(\beta)$ stabilizes with L for all values of β plotted. Comparing Figure 2a and Figure 2b shows that the value of r after which the tail $E_{L,T}(r)$ becomes smaller than a fixed value ϵ is more uniform for $T = 10$ than for $T = 2/3$. And in either case, the tail decays exponentially with r (justifying the idealization that $d_\gamma = 0$ for $\mathrm{ord}(\gamma) > r$). These results show that Assumption 5.1 is a good assumption at higher temperatures. The interpretation is that for an Ising model at a higher temperature, the decay of correlations is more pronounced.

We observe a similar effect in Figure 3 for the GTR model, but with η replacing temperature T. Here, η is to be interpreted as a measure of rate of mixing for the markov process. At high values of η, Assumption 5.1 is a good assumption for the GTR model.

9 Conclusion

In this paper, we studied approximation and estimation error bounds for Transformer based Masked Diffusion Models trained with continuous time NELBO loss when the data has low complexity structure. This is, to our knowledge, the first such study in this setting. As a result, we obtain $O\big(M\big((\log m)^{r+1} + (M+2)^2(\log m)^3\big)\big)$ dependence of the optimal network size. Extending our analysis to more general discrete distributions with weaker assumptions remains a direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to provide critical ingredients for proving mathematical claims, assist in the writing of proofs, provide feedback on experiments and implement methods. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual framework, formulate mathematical claims, propose or refine hypotheses, support qualitative and thematic data analysis, interpret results or reformat datasets. Translation is not applicable to this work. Additionally, we used generative AI tools to create scientific figures, suggest experimental parameters, draft parts of the paper, summarize existing literature, brainstorming, search for information, edit the paper to improve readability, identify relevant literature and format references.

We provide details on how AI was used for the mathematical results. GPT 5.5 and GPT 5.6 Sol were used to fetch and understand the literature cited in this paper. Especially, GPT 5.5 helped us understand the proofs of Takakura & Suzuki (2023), following which we developed the proof of Theorem 4.1. The proof of Theorem C.1 using CTMC theory was written collaboratively with these models, and is essentially same as the corresponding proof for the discrete score which was first given in Lou et al. (2024) (we use posterior instead of the score). As such, we do not claim it to be a new result.

The experimental setup in this paper was developed by the authors, and improved with the feedback of GPT 5.6 Sol. The model was also used to choose the parameters for the experiments. The code used to run the experiments was written using GPT 5.6 Sol, with the experimental setup provided as prompt.

We have reviewed all AI-assisted work. We have checked the steps written by GPT in the proof of Theorem C.1, and identified other papers in literature which state it for the discrete score. The code for the experiments was manually reviewed by the authors. 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 is primarily theoretical and does not involve human subjects, personally identifiable information, or other sensitive data. The experiments use only publicly available datasets. We are not aware of any additional ethical concerns arising from our research.

REPRODUCIBILITY STATEMENT

In this statement, we discuss the efforts that have been made to ensure reproducibility. For the theoretical results, all assumptions are stated explicitly in the main text, and complete proofs are provided in the appendix. For the numerical experiments, the full experimental setup is described in Section 7 and Appendix H, including the Ising and GTR distributions, all parameter choices, the sequence lengths considered, the computation of the posterior logits and their coefficient expansions, and the diagnostics used in the figures. These experiments are designed to be reproducible directly from the mathematical specification given there.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html. 1, 3

Joe Benton, Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. From denoising diffusions to denoising markov models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):286–301, 2024. doi: 10.1093/jrsssb/qkae005. URL https://academic.oup.com/jrsssb/article/86/2/286/7564909. 1

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In

Advances in Neural Information Processing Systems, volume 35, pp. 28266–28279, 2022. doi: 10.52202/068431-2049. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b5b528767aa35f5b1a60fe0aaeca0563-Abstract-Conference.html. 1, 3, 24

Hongrui Chen and Lexing Ying. Convergence analysis of discrete diffusion model: Exact implementation through uniformization. *Journal of Machine Learning*, 4(2):108–127, 2025. doi: 10.4208/jml.240812. URL https://www.global-sci.com/jml/article/view/13211. 1

Sitan Chen, Kevin Cong, and Jerry Li. Optimal inference schedules for masked diffusion models. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1279–1311. PMLR, 2026. URL https://proceedings.mlr.press/v336/chen26e.html. 1

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical Optimal Transport: École d'Été de Probabilités de Saint-Flour XLIX – 2019*, volume 2364 of *Lecture Notes in Mathematics*. Springer, 2025. doi: 10.1007/978-3-031-85160-5. URL https://link.springer.com/book/10.1007/978-3-031-85160-5. Proposition 2.6, p. 55 in the freely available preprint: https://arxiv.org/abs/2407.18163. 28

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. DiffusionBERT: Improving generative masked language models with diffusion models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4521–4534. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.248. URL https://aclanthology.org/2023.acl-long.248/. 1

Xunpeng Huang, Yingyu Lin, Nishant Jain, Kaibo Wang, Difan Zou, Yian Ma, and Tong Zhang. On the ϵ-free inference complexity of absorbing discrete diffusion, 2025. URL https://arxiv.org/abs/2509.21835. 1

Moongyu Jeon, Sangwoo Shin, Dongjae Jeon, and Albert No. Information-theoretic discrete diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 20913–20939, 2025. doi: 10.52202/085713-0620. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1a9b888813ffc9349371087c95c22af3-Abstract-Conference.html. 24

Gen Li and Changxiao Cai. Breaking AR's sampling bottleneck: Provable acceleration via diffusion language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0392. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/111298628bfb153ee1f84b10fec3a8b9-Abstract-Conference.html. 1

Yuchen Liang, Renxiang Huang, Lifeng Lai, Ness Shroff, and Yingbin Liang. Absorb and converge: Provable convergence guarantee for absorbing discrete diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-0685. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1d571ea833394630c1dec71664f16cd9-Abstract-Conference.html. 1

Yuchen Liang, Yingbin Liang, Lifeng Lai, and Ness Shroff. Discrete diffusion models: Novel analysis and new sampler guarantees. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-5517. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f1e7c90552850afcc2558d78950c519d-Abstract-Conference.html. 1

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR, 2024. URL https://proceedings.mlr.press/v235/lou24a.html. 1, 10, 24

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1689. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/48b383b24230e0e6e649d9c98dae4d8c-Abstract-Conference.html. 1

Yuto Nishimura and Taiji Suzuki. Minimax optimality of convolutional neural networks for infinite dimensional input-output problems and separation from kernel methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4e11ee35e17c9807b1a9616b717e548e-Paper-Conference.pdf. 20

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*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a365e37c18fb91af547a2f0012a89e98-Abstract-Conference.html. 1

Le-Tuyet-Nhi Pham, Dario Shariatian, Antonio Ocello, Giovanni Conforti, and Alain Oliviero Durmus. Discrete markov probabilistic models: An improved discrete score-based framework with sharp convergence bounds under minimal assumptions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49195–49258. PMLR, 2025. URL https://proceedings.mlr.press/v267/pham25a.html. 1

Yinuo Ren, Haoxuan Chen, Grant M. Rotskoff, and Lexing Ying. How discrete and continuous diffusion meet: Comprehensive analysis of discrete diffusion models via a stochastic integral framework. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6a9305d8e1dc254308a2c2e918108007-Abstract-Conference.html. 1, 24

Yinuo Ren, Haoxuan Chen, Yuchen Zhu, Wei Guo, Yongxin Chen, Grant M. Rotskoff, Molei Tao, and Lexing Ying. Fast solvers for discrete diffusion models: Theory and applications of high-order algorithms. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-5575. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f46ddea413df86832418c5e04e59644f-Abstract-Conference.html. 1

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*, volume 37, pp. 130136–130184, 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb0b13cc515724ab8015bc978fdde0ad-Abstract-Conference.html. 1, 4, 25

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103131–103167, 2024. doi: 10.52202/079017-3277. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bad233b9849f019aead5e5cc60cef70f-Abstract-Conference.html. 1

Aadithya Srikanth, Mudit Gaur, and Vaneet Aggarwal. Discrete state diffusion models: A sample complexity perspective. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3547–3555. PMLR, 2026. URL https://proceedings.mlr.press/v300/srikanth26a.html. 1, 2

Maojiang Su, Mingcheng Lu, Jerry Yao-Chieh Hu, Shang Wu, Zhao Song, Alex Reneau, and Han Liu. A theoretical analysis of discrete flow matching generative models, 2025. URL https://arxiv.org/pdf/2509.22623. 2

Danica J. Sutherland. CPSC 532D: Statistical Learning Theory—5. Rademacher Complexity, 2026. URL https://www.cs.ubc.ca/~dsuth/532D/26w1/notes/5-rademacher.pdf. University of British Columbia lecture notes, Fall 2026. Theorem 5.7, equations (5.7)–(5.8), p. 9. 28

Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of Transformers for sequence-to-sequence functions with infinite dimensional input. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33416–33447, 2023. Theorem 5.3 and Appendix H. PMLR page: https://proceedings.mlr.press/v202/takakura23a.html; public PDF: https://proceedings.mlr.press/v202/takakura23a/takakura23a.pdf. 3, 5, 6, 10, 14, 15, 17, 19, 20, 29

Shintaro Wakasugi and Taiji Suzuki. State size independent statistical error bound for discrete diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 154062–154097, 2025. doi: 10.52202/085713-4637. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/caabce23491e5418f4c9ea348359fcd2-Abstract-Conference.html. 1, 2

Zhengyan Wan, Yidong Ouyang, Qiang Yao, Liyan Xie, Fang Fang, Hongyuan Zha, and Guang Cheng. Error analysis of discrete flow with generator matching. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://openreview.net/forum?id=tiPoSBaOiY. 2

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal large diffusion language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4636. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/caa934a507a952698d54efb24845fc4b-Abstract-Conference.html. 1

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017. doi: 10.1016/j.neunet.2017.07.002. URL https://arxiv.org/pdf/1610.01145. 20

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models, 2025. URL https://arxiv.org/abs/2508.15487. 1

Zikun Zhang, Zixiang Chen, and Quanquan Gu. Convergence of score-based discrete diffusion models: A discrete-time analysis. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/560b4a3ad23137d641258d8807de13e9-Abstract-Conference.html. 1

Zixuan Zhang, Hengyu Fu, Zhuoran Yang, Mengdi Wang, Tuo Zhao, and Minshuo Chen. Generalization bounds for discrete diffusion: Statistical advantage of masking. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://openreview.net/forum?id=Ofhq7nBVHu. 1, 2

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1898. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/52190a0362148d179f1cbd9080956872-Abstract-Conference.html. 1

Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. A reparameterized discrete diffusion model for text generation. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=PEQFHRUFca. 1

A Additional definitions and corollaries

The following material is not needed to parse the main theorem statements, but is included here for completeness.

A.1 Masking-time construction of the forward process

A convenient way to construct the forward process is to assign each token an independent masking time ξ_i, independent of X, with

$$\mathbb{P}(\xi_i > t) = \alpha_t, \qquad \alpha_t := \alpha(t). \tag{A.1}$$

We then define the forward process Z_t as

$$(Z_t)_i = \begin{cases} X_i, & \xi_i > t, \\ \mathtt{m}, & \xi_i \le t, \end{cases} \qquad i \in [L]. \tag{A.2}$$

Thus a token is clean until its masking time and remains masked thereafter. Let $q_t(z \mid x) := \mathbb{P}(Z_t = z \mid X = x)$. Since the masking times $\xi_1, \dots, \xi_L$ are independent, conditional on $X = x$ the coordinates of Z_t are independent. Moreover,

$$\mathbb{P}((Z_t)_i = x_i \mid X = x) = \mathbb{P}(\xi_i > t) = \alpha_t, \qquad \mathbb{P}((Z_t)_i = \mathtt{m} \mid X = x) = 1 - \alpha_t.$$

Consequently,

$$q_t(z \mid x) = \prod_{j=1}^{L} \left(\alpha_t \mathbf{1}_{\{z_j = x_j\}} + (1-\alpha_t)\mathbf{1}_{\{z_j = \mathtt{m}\}}\right), \qquad z \in \mathcal{E}^L, \quad x \in \mathcal{A}^L. \tag{A.3}$$

Hence, at each time t, each coordinate is retained independently with probability α_t and otherwise replaced by $\mathtt{m}$. The corresponding forward marginal is

$$q_t(z) = \sum_{x \in \mathcal{A}^L} P(x) q_t(z \mid x).$$

A.2 Detailed one-block Transformer formulae

Formulae. For $Z = (z_j)_{j\in\mathbb{Z}} \in (\mathbb{R}^D)^{\mathbb{Z}}$, put $Z_{j,U} := [z_{j-U}, \ldots, z_{j+U}] \in \mathbb{R}^{D\times(2U+1)}$. A sliding-window attention layer, with $W_h^Q, W_h^K \in \mathbb{R}^{D_h^{\mathrm{qk}}\times D}$ and $W_h^V \in \mathbb{R}^{D\times D}$, is

$$\mathsf{Attn}_{U,H}(Z)_j := z_j + \sum_{h=1}^{H} W_h^V Z_{j,U} \operatorname{Softmax}\Big((W_h^K Z_{j,U})^\top W_h^Q z_j\Big).$$

A position-wise ReLU FNN is $(\mathsf{FNN}_g Z)_j := g(z_j)$, where, for affine maps $\mathcal{A}_\ell : \mathbb{R}^{D_{\ell-1}^{\mathrm{FF}}} \to \mathbb{R}^{D_\ell^{\mathrm{FF}}}$, $\mathcal{A}_\ell(z) := A_\ell^{\mathrm{FF}} z + c_\ell^{\mathrm{FF}}$,

$$g := \mathcal{A}_{L_{\mathrm{net}}} \circ \mathrm{ReLU} \circ \mathcal{A}_{L_{\mathrm{net}}-1} \circ \cdots \circ \mathrm{ReLU} \circ \mathcal{A}_1 : \mathbb{R}^D \to \mathbb{R},$$

with $D_0^{\mathrm{FF}} = D$, $D_{L_{\mathrm{net}}}^{\mathrm{FF}} = 1$, and $\max_{\ell < L_{\mathrm{net}}} D_\ell^{\mathrm{FF}} \le W_{\mathrm{net}}$. A one-block sliding-window Transformer is the composition

$$\widehat{F}_j(X) := \big(\mathsf{FNN}_g \circ \mathsf{Attn}_{U,H} \circ \mathrm{Enc}_{\mathrm{PE}}\big)(X)_j,$$
$$(\mathrm{Enc}_{\mathrm{PE}}\, X)_j := W_{\mathrm{emb}} e_{X_j} + \mathrm{pe}_j, \qquad \mathrm{PE} = (\mathrm{pe}_j)_{j\in\mathbb{Z}}.$$

Here W_{emb} is the token-embedding matrix, PE is the positional encoding, U, D, H are the half-window, embedding dimension, and number of heads, D_h^{qk} is the query/key dimension of head h, and $L_{\mathrm{net}}, W_{\mathrm{net}}$ are the FNN depth and maximal hidden width. Define

$$S_{\mathrm{net}} := \sum_{\ell=1}^{L_{\mathrm{net}}} \big(\|A_\ell^{\mathrm{FF}}\|_0 + \|c_\ell^{\mathrm{FF}}\|_0\big),$$
$$B_{\mathrm{net}} := \max\{|\theta| : \theta \text{ is a scalar network parameter}\}, \qquad B_{\mathrm{cov}} := \max\{1, B_{\mathrm{net}}, \|\mathrm{PE}\|_\infty\}.$$

A.3 Transformer covering number

A.3.1 Covering number

In our notation, Takakura & Suzuki (2023, Theorem 5.3) gives

$$\log N(\eta, \mathcal{T}, \|\cdot\|_\infty) \lesssim N_{\mathrm{blk}}^3 L_{\mathrm{net}} (S_{\mathrm{net}} + HD^2) \log\left(\frac{DHL_{\mathrm{net}} W_{\mathrm{net}} B_{\mathrm{cov}}}{\eta}\right), \tag{A.4}$$

where $\mathcal{T}$ denotes the class of N_{blk}-block sliding window Transformers with parameters $(U, D, H, L_{\mathrm{net}}, W_{\mathrm{net}}, S_{\mathrm{net}}, B_{\mathrm{cov}}; \mathrm{PE})$.

A.4 Regularity-class details

Fix $\beta > 0$, $1 \le \tau \le \infty$, and $r \in \mathbb{N}$. Define

$$\|f\|_{\mathcal{B}_{\beta,\tau}^{(r)}} := \left\| \left(2^{\beta\varrho(\gamma)} |d_\gamma|\right)_{\gamma\in\mathfrak{A}} \right\|_{\ell^\tau},$$

with the additional condition $d_\gamma = 0$ if $\mathrm{ord}(\gamma) > r$.

We consider the class of functions

$$\mathcal{B}^{(r)}_{\beta,\tau} := \Big\{ f = \sum_{\gamma\in\mathfrak{A}} d_\gamma b_\gamma : \|f\|_{\mathcal{B}^{(r)}_{\beta,\tau}} < \infty,\ d_\gamma = 0 \text{ if } \operatorname{ord}(\gamma) > r \Big\}.$$

This space is not closed under the uniform topology (in which we develop the approximation theory), but its unit ball $U(\mathcal{B}^{(r)}_{\beta,\tau}) := \Big\{ f \in \mathcal{B}^{(r)}_{\beta,\tau} : \|f\|_{\mathcal{B}^{(r)}_{\beta,\tau}} \le 1 \Big\}$ is closed.

Interpretation The case $\tau = 1$ has a simple interpretation in terms of metric spaces. For $x, y \in E^{\mathbb{Z}}$, let

$$\operatorname{sep}(x,y) := \min\{|j| : x_j \neq y_j\}, \qquad d_\beta(x,y) := 2^{-\beta \operatorname{sep}(x,y)},$$

with $\operatorname{sep}(x,x) = \infty$. Thus d_β is the bidirectional analogue of the usual shift-space metric: two sequences are $2^{-\beta k}$ close (for $k \ge 1$) when they agree on indices $-k+1, \dots 0, \dots k-1$.

Lemma A.1. *Let*

$$f = \sum_{\gamma\in\mathfrak{A}} d_\gamma b_\gamma \in \mathcal{B}^{(r)}_{\beta,1}.$$

Then f is Lipschitz with respect to d_β, with

$$|f(x) - f(y)| \le 2\|f\|_{\mathcal{B}^{(r)}_{\beta,1}}\, d_\beta(x,y).$$

Proof. Let $s = \operatorname{sep}(x,y)$. If $\varrho(\gamma) < s$, then $b_\gamma(x) = b_\gamma(y)$, since x and y agree on every coordinate used by b_γ. Since $\|b_\gamma\|_\infty \le 1$,

$$|f(x) - f(y)| \le 2 \sum_{\varrho(\gamma)\ge s} |d_\gamma| \le 2\, 2^{-\beta s} \sum_{\gamma} 2^{\beta\varrho(\gamma)} |d_\gamma| = 2\, d_\beta(x,y)\, \|f\|_{\mathcal{B}^{(r)}_{\beta,1}}.$$

□

A.5 Consequences of the approximation theorem

Corollary A.2 (Bound in terms of network size)**.** *Let $a_{\mathrm{par}} := \max\{r+1, 3\}$. For all sufficiently large N, there is a Transformer with at most N parameters such that*

$$\sup_{j\in\mathbb{Z}} \|\widehat{F}_{N,j} - F^\circ_j\|_\infty \le C N^{\kappa/a_{\mathrm{par}}} \exp\Big(-cN^{1/a_{\mathrm{par}}}\Big).$$

The same bound holds in $L^2(\mu)$.

We use this result to bound the number of parameters required to achieve a certain approximation error. Let $N(\varepsilon)$ be the minimum number of parameters so that $\sup_{j\in\mathbb{Z}} \|\widehat{F}_{N,j} - F^\circ_j\|_\infty \le \epsilon$.

Corollary A.3 (Parameter count for error atmost ε)**.** *It is sufficient to use $K = \frac{1}{\beta}\log_2\frac{1}{\varepsilon} + O\left(\log\log\frac{1}{\varepsilon}\right)$, and therefore*

$$N(\varepsilon) \le C\left(\log\frac{1}{\varepsilon}\right)^{\max\{r+1,3\}}.$$

A.6 Restriction used in the covering argument

be the corresponding K-dependent scalar class with fixed positional encoding. We use the uniform norm from Takakura & Suzuki (2023, Theorem 5.3); restriction to padded MDLM inputs and scalar-coordinate projection are 1-Lipschitz, so

$$\sup_{\substack{z\in\mathcal{Z}_P \\ i\in\mathcal{M}(z)}} |F_i(\iota z) - G_i(\iota z)| \le \|F - G\|_{\infty,\mathrm{TS}}. \tag{A.5}$$

A.7 Conditional loss notation

For $p \in \mathcal{P}_{K,\lambda}^{\mathrm{bd}}$ and $x \in \mathrm{supp}(P)$, the loss on a clean sequence is

$$J_p(x) := \int_0^1 w(t) \sum_{z \in \mathcal{E}^L} q_t(z \mid x) \sum_{i \in \mathcal{M}(z)} -\log p_i(x_i \mid z)\, dt, \tag{A.6}$$

Thus $R(p) = \mathbb{E} J_p(X)$. In Appendix A.8, we use the class

$$\mathscr{L}_{K,\lambda} := \{J_p : p \in \mathcal{P}_{K,\lambda}^{\mathrm{bd}}\}$$

and apply empirical risk minimization theory to bound the estimation error.

A.8 Corollaries of the MDLM theorems

Corollary A.4 (Error Decomposition)**.** *Fix a predictor class $\mathcal{P}_K$, and let $\widehat{p}_{K,m} \in \mathcal{P}_K$ be any data-dependent estimator. Then, whenever Theorem 6.1 applies to $\widehat{p}_{K,m}$,*

$$D_{\mathrm{KL}}\left(P \middle\| P^{(\infty)}_{\widehat{p}_{K,m}}\right) \le \mathrm{Approx}(\mathcal{P}_K) + \mathrm{Est}_{K,m},$$
$$D_{\mathrm{KL}}\left(P \middle\| P^{(T)}_{\widehat{p}_{K,m}}\right) \le \mathrm{Approx}(\mathcal{P}_K) + \mathrm{Est}_{K,m} + \mathrm{Disc}_{\widehat{p}_{K,m}}(T).$$

Corollary A.5. *Under the conditions of Theorem 6.3, there is a constant $C > 0$ depending only on the fixed constants in the approximation and covering theorems such that, defining*

$$\overline{V}_K = M\left\{(K+1)^{r+2} + Q^2(K+1)^4\right\}, \qquad \overline{C}_K = CQ(1 \vee R_\lambda)(K+1)^{r+6}\log(K+2),$$

we have, with probability at least $1-\delta$,

$$R(\widehat{p}_m) - H(P) \le L\log\frac{1}{1-\lambda} + 2L\rho_K + C\mathcal{B}_{K,\lambda}\sqrt{\frac{\overline{V}_K}{m}} \times \left[1 + \sqrt{\log\left(\overline{C}_K\left(1 + \frac{8L}{\mathcal{B}_{K,\lambda}}\right)\right)}\right] + \mathcal{B}_{K,\lambda}\sqrt{\frac{2}{m}\log\frac{2}{\delta}}.$$

A.9 Tradeoff between Approximation and Finite Sample Statistical errors

Corollary A.5 exhibits an explicit tradeoff in the parameter K. Increasing K enlarges the Transformer architecture and decreases the approximation contribution $2L\rho_K$, but simultaneously increases the covering-number contribution to the finite sample statistical error. The two terms have qualitatively different dependence on K: the former decays exponentially up to a polynomial factor, whereas the latter grows only polynomially, up to logarithmic factors.

To make this comparison precise, throughout this subsection fix $\lambda \in (0,1)$ and regard $M, r, \beta, \tau, R_\lambda, \delta$ and the constants in the approximation and covering bounds as fixed. Set

$$a := \max\{r+2, 4\}, \qquad c_\beta := \beta\log 2. \tag{A.7}$$

Define the K-dependent approximation contribution

$$\mathfrak{A}_K := 2L\rho_K = 2LC_{\mathrm{tr}}R_\lambda(K+1)^\kappa e^{-c_\beta K}, \tag{A.8}$$

and the complexity-dependent part of the finite sample statistical error bound in Corollary A.5 by

$$\mathfrak{E}_{K,m} := C\mathcal{B}_{K,\lambda}\sqrt{\frac{\overline{V}_K}{m}}\left[1 + \sqrt{\log\left(\overline{C}_K\left(1 + \frac{8L}{\mathcal{B}_{K,\lambda}}\right)\right)}\right]. \tag{A.9}$$

Proposition A.6 (Optimal K)**.** *Under the conditions of Corollary A.5, suppose that the structural parameters M, r, β, τ, ν and R_λ are bounded uniformly in L. Then*

$$\frac{\mathfrak{A}_K}{L} = O\big(K^\kappa e^{-c_\beta K}\big), \qquad \frac{\mathfrak{E}_{K,m}}{L} = \Theta\left(\frac{K^{a/2}\sqrt{\log K}}{\sqrt{m}}\right), \qquad K \to \infty,$$

where the implicit constants are independent of L, K, and m. Consequently, if $R_\lambda > 0$, balancing the two terms gives the optimal K as

$$K_m = \frac{\log m}{2\beta\log 2} + O(\log\log m),$$

again independently of L.

B Proof of Theorem 4.1 and Its Consequences

This appendix contains the proof of the approximation theorem for the sliding window transformer and of corollary. The proof largely follows Takakura & Suzuki (2023).

Set $\varepsilon_K := (K+1)^\kappa 2^{-\beta K}$. Although $F_j^\circ(X) = f(\Sigma_j X)$ may depend on the entire sequence, the regularity assumption on f implies that interactions involving coordinates far from the origin matter less. Hence we define

$$\mathfrak{A}_{K,r} := \{\gamma \in \mathfrak{A} : \varrho(\gamma) \le K,\ \mathrm{ord}(\gamma) \le r\}, \qquad f_K := \sum_{\gamma \in \mathfrak{A}_{K,r}} d_\gamma b_\gamma.$$

Every term retained in f_K depends on at most r coordinates, and all of these coordinates lie in $[-K, K]$. Hence $f_K(\Sigma_j X)$ depends only on $X_{j-K}, \dots, X_{j+K}$.

Bounding the truncation error $\|f - f_K\|_\infty$ using regularity assumptions. First, $G_r(K) := |\mathfrak{A}_{K,r}| = \sum_{\ell=0}^r \binom{2K+1}{\ell}(q-1)^\ell \le C_{q,r}(K+1)^r$. Hence, if $n_r(s) := |\{\gamma : \varrho(\gamma) = s,\ \mathrm{ord}(\gamma) \le r\}|$, then $n_r(s) = G_r(s) - G_r(s-1) \le C_{q,r}(s+1)^{r-1}$. Since $\|b_\gamma\|_\infty \le 1$, we have $\|f - f_K\|_\infty \le \sum_{\varrho(\gamma)>K} |d_\gamma|$. If $\tau = 1$, then

$$\sum_{\varrho(\gamma)>K} |d_\gamma| \le 2^{-\beta K} \sum_\gamma 2^{\beta\varrho(\gamma)} |d_\gamma| \le 2^{-\beta K}.$$

If $\tau > 1$, Hölder's inequality and (2) give

$$\sum_{\varrho(\gamma)>K} |d_\gamma| \le \left(\sum_{\varrho(\gamma)>K} 2^{-\beta\tau'\varrho(\gamma)}\right)^{1/\tau'} \|f\|_{\mathcal{B}^{(r)}_{\beta,\tau}} \le C \left(\sum_{s>K} (s+1)^{r-1} 2^{-\beta\tau' s}\right)^{1/\tau'} \le C(K+1)^{(r-1)/\tau'} 2^{-\beta K}.$$

Therefore

$$\boxed{\|f - f_K\|_\infty \le C(K+1)^\kappa 2^{-\beta K}} \tag{B.1}$$

So now we focus on approximating $f_K(\Sigma_j X)$. First, write it as the composition

$$X \longmapsto \Sigma_j X \longmapsto (e_{X_{j-K}}, \dots, e_{X_{j+K}}) \longmapsto f_K(\Sigma_j X)$$

Introduce the map

$$\Gamma_K(X) : X \mapsto (e_{X_{-K}}, \dots, e_{X_K}) \in \mathbb{R}^{q(2K+1)}$$

Then we have the diagram

$$X \longmapsto \Sigma_j X \xmapsto{\Gamma_K} (e_{X_{j-K}}, \dots, e_{X_{j+K}}) \longmapsto f_K(\Sigma_j X).$$

Following the construction of Takakura & Suzuki (2023), we approximate the map Γ_K using Attention.

Construction of Attention block. Fix $h \in \{1, \dots, H_K\}$, where $H_K = 2K+1$ and $t_h = -K + h - 1$. Head h is constructed to retrieve the one-hot vector $e_{X_{j+t_h}}$ while processing position j.

1. Token embedding. Let the alphabet be $[q] = \{1, \dots, q\}$, and represent $X_j \in [q]$ by $e_{X_j} \in \mathbb{R}^q$, where $(e_{X_j})_a = \mathbf{1}\{X_j = a\}$. Set $D_K = q + qH_K + 2$ and $\phi = 2\pi/(2K+1)$, and define

$$z_j^{(0)} = \begin{bmatrix} e_{X_j} \\ 0_{qH_K} \\ \cos(j\phi) \\ \sin(j\phi) \end{bmatrix} \in \mathbb{R}^{D_K}.$$

The middle qH_K coordinates are reserved for storing the H_K fetched one-hot vectors.

2. Key map. Here $D_h^{\mathrm{qk}} = 2$. The key matrix is

$$W_h^K = \begin{bmatrix} 0_{1\times(D_K-2)} & 1 & 0 \\ 0_{1\times(D_K-2)} & 0 & 1 \end{bmatrix} \in \mathbb{R}^{2\times D_K}.$$

Thus, for a candidate position $j+s$,

$$k_{j+s}^{(h)} := W_h^K z_{j+s}^{(0)} = \begin{bmatrix} \cos((j+s)\phi) \\ \sin((j+s)\phi) \end{bmatrix}.$$

3. Query map. For head h, define

$$W_h^Q = \chi \begin{bmatrix} 0_{1\times(D_K-2)} & \cos(t_h\phi) & -\sin(t_h\phi) \\ 0_{1\times(D_K-2)} & \sin(t_h\phi) & \cos(t_h\phi) \end{bmatrix} \in \mathbb{R}^{2\times D_K}.$$

Therefore

$$\begin{aligned} q_j^{(h)} := W_h^Q z_j^{(0)} &= \chi \begin{bmatrix} \cos(t_h\phi) & -\sin(t_h\phi) \\ \sin(t_h\phi) & \cos(t_h\phi) \end{bmatrix} \begin{bmatrix} \cos(j\phi) \\ \sin(j\phi) \end{bmatrix} \\ &= \chi \begin{bmatrix} \cos((j+t_h)\phi) \\ \sin((j+t_h)\phi) \end{bmatrix}. \end{aligned}$$

4. Attention score. For $s \in [-K:K]$, the score assigned to position $j+s$ is $S_{j,s}^{(h)} = (k_{j+s}^{(h)})^\top q_j^{(h)}$. Hence

$$\begin{aligned} S_{j,s}^{(h)} &= \chi \begin{bmatrix} \cos((j+s)\phi) & \sin((j+s)\phi) \end{bmatrix} \begin{bmatrix} \cos((j+t_h)\phi) \\ \sin((j+t_h)\phi) \end{bmatrix} \\ &= \chi \cos((s-t_h)\phi). \end{aligned}$$

In particular, $S_{j,t_h}^{(h)} = \chi$, whereas $S_{j,s}^{(h)} < \chi$ for $s \neq t_h$. Thus the unique maximizer is the candidate at position $j+t_h$.

5. Softmax weights. Define

$$\alpha_{j,s}^{(h)} := \frac{\exp(S_{j,s}^{(h)})}{\sum_{r=-K}^{K} \exp(S_{j,r}^{(h)})} = \frac{\exp(\chi\cos((s-t_h)\phi))}{\sum_{r=-K}^{K} \exp(\chi\cos((r-t_h)\phi))}.$$

For large χ, $\alpha_{j,t_h}^{(h)} \approx 1$ and $\alpha_{j,s}^{(h)} \approx 0$ for $s \neq t_h$, so the softmax approximates the hard choice $s = t_h$.

6. Value map. Let $J_h : \mathbb{R}^q \to \mathbb{R}^{D_K}$ inject a vector into the h-th reserved block:

$$J_h x = \begin{bmatrix} 0_q \\ 0_{q(h-1)} \\ x \\ 0_{q(H_K-h)} \\ 0 \\ 0 \end{bmatrix}.$$

Let $P_{\mathrm{tok}} = \begin{bmatrix} I_q & 0_{q\times(qH_K+2)} \end{bmatrix}$, so that $P_{\mathrm{tok}} z_j^{(0)} = e_{X_j}$, and define $W_h^V := J_h P_{\mathrm{tok}}$. Then

$$W_h^V z_{j+s}^{(0)} = J_h e_{X_{j+s}},$$

so the value map copies the entire one-hot token into the h-th reserved block.

7. Output of head h. The contribution of head h at position j is

$$u_j^{(h)} := \sum_{s=-K}^{K} \alpha_{j,s}^{(h)} W_h^V z_{j+s}^{(0)} = J_h \sum_{s=-K}^{K} \alpha_{j,s}^{(h)} e_{X_{j+s}}.$$

Since the attention weights concentrate at $s = t_h$,

$$u_j^{(h)} \approx J_h e_{X_{j+t_h}}.$$

Equivalently, inside the h-th reserved block, $\sum_{s=-K}^{K} \alpha_{j,s}^{(h)} e_{X_{j+s}} \approx e_{X_{j+t_h}}$. Thus head h retrieves $e_{X_{j+t_h}}$ while processing position j.

8. Combining all heads. With all $H_K = 2K + 1$ heads and the residual connection,

$$z_j^{(1)} = z_j^{(0)} + \sum_{h=1}^{H_K} u_j^{(h)} \approx z_j^{(0)} + \sum_{h=1}^{H_K} J_h e_{X_{j+t_h}}.$$

Since $(t_1, \ldots, t_{H_K}) = (-K, \ldots, K)$, the reserved coordinates contain approximately

$$\begin{bmatrix} e_{X_{j-K}} \\ e_{X_{j-K+1}} \\ \vdots \\ e_{X_{j+K}} \end{bmatrix}.$$

Hence, if $\Pi_{\text{win}} : \mathbb{R}^{D_K} \to \mathbb{R}^{qH_K}$ extracts the reserved coordinates, then

$$\Pi_{\text{win}} z_j^{(1)} \approx \left(e_{X_{j-K}}, \ldots, e_{X_{j+K}}\right) = \Gamma_K(\Sigma_j X).$$

So we get the diagram

$$X \longmapsto \Sigma_j X \xrightarrow{\Gamma_K} (e_{X_{j-K}}, \ldots, e_{X_{j+K}}) \longmapsto f_K(\Sigma_j X).$$
$$\Pi_{\text{win}} z_j^{(1)} \approx$$

Bounding the error $\|\Pi_{\text{win}} z_j^{(1)} - \Gamma_K(\Sigma_j X)\|_\infty$

1. Score gap. For head h, the score of the candidate at relative offset $s \in [-K : K]$ is $S_{j,s}^{(h)} = \chi \cos((s - t_h)\phi)$, where $\phi = 2\pi/(2K + 1)$. At the target offset, $S_{j,t_h}^{(h)} = \chi$. If $s \neq t_h$, then $s - t_h$ is a nonzero residue modulo $2K + 1$, and therefore $\cos((s - t_h)\phi) \leq \cos(2\pi/(2K + 1))$. Hence

$$\begin{aligned} S_{j,t_h}^{(h)} - S_{j,s}^{(h)} &\geq \chi \left(1 - \cos \frac{2\pi}{2K+1}\right) \\ &= 2\chi \sin^2 \left(\frac{\pi}{2K+1}\right) \geq \frac{c\chi}{K^2}, \end{aligned}$$

for an absolute constant $c > 0$. Thus $S_{j,t_h}^{(h)} - S_{j,s}^{(h)} \geq c\chi/K^2$ for every $s \neq t_h$. This is the same argument used in Takakura & Suzuki (2023), Appendix E, adapted to the present choice $\phi = 2\pi/(2K + 1)$.

2. Softmax concentration. Takakura & Suzuki (2023), Lemma C.1, states that if $\theta \in \mathbb{R}^d$ has an index i_* such that $\theta_{i_*} \geq \theta_i + \delta$ for every $i \neq i_*$, then

$$\| \operatorname{Softmax}(\theta) - e_{i_*} \|_1 \leq 2de^{-\delta}.$$

Apply this with $d = H_K = 2K + 1$, $i_* = t_h$, and $\delta = c\chi/K^2$. If $a_j^{(h)}$ denotes the attention-weight vector of head h at position j, then

$$\begin{aligned} \|a_j^{(h)} - e_{t_h}\|_1 &\leq 2(2K + 1) \exp\left(-\frac{c\chi}{K^2}\right) \\ &\leq CK \exp\left(-\frac{c\chi}{K^2}\right), \end{aligned}$$

for an absolute constant $C > 0$.

3. Error in the value retrieved by one head. The actual contribution of head h is $u_j^{(h)} = J_h \sum_{s=-K}^{K} \alpha_{j,s}^{(h)} e_{X_{j+s}}$, whereas the ideal contribution is $\widetilde{u}_j^{(h)} = J_h e_{X_{j+t_h}}$. Hence

$$u_j^{(h)} - \widetilde{u}_j^{(h)} = J_h \sum_{s=-K}^{K} \left(\alpha_{j,s}^{(h)} - \mathbf{1}\{s = t_h\}\right) e_{X_{j+s}}.$$

Since J_h is a coordinate injection and $\|e_{X_{j+s}}\|_\infty = 1$,

$$\begin{aligned}\|u_j^{(h)} - \widetilde{u}_j^{(h)}\|_\infty &\le \sum_{s=-K}^{K} \left|\alpha_{j,s}^{(h)} - \mathbf{1}\{s = t_h\}\right| \\ &= \|a_j^{(h)} - e_{t_h}\|_1 \\ &\le CK \exp\left(-\frac{c\chi}{K^2}\right).\end{aligned}$$

This is the block-valued analogue of the value-error estimate in Takakura & Suzuki (2023), Appendix E.

4. Combine the heads. Note that the combined output of attention heads approximates the quantity $\widetilde{z}_j := \sum_{h=1}^{H_K} \widetilde{u}_j^{(h)}$. The actual and ideal attention outputs satisfy $z_j^{(1)} - \widetilde{z}_j = \sum_{h=1}^{H_K}(u_j^{(h)} - \widetilde{u}_j^{(h)})$. The image of J_h is the h-th reserved block, and these blocks are pairwise disjoint. Therefore

$$\left\|\sum_{h=1}^{H_K}\left(u_j^{(h)} - \widetilde{u}_j^{(h)}\right)\right\|_\infty = \max_{1\le h\le H_K} \|u_j^{(h)} - \widetilde{u}_j^{(h)}\|_\infty.$$

Consequently,

$$\|z_j^{(1)} - \widetilde{z}_j\|_\infty \le CK \exp\left(-\frac{c\chi}{K^2}\right).$$

5. Extract the window. Let $\Pi_{\text{win}} : \mathbb{R}^{D_K} \to \mathbb{R}^{qH_K}$ extract the qH_K reserved coordinates. Since Π_{win} is a coordinate projection, $\|\Pi_{\text{win}}x - \Pi_{\text{win}}y\|_\infty \le \|x - y\|_\infty$. Hence

$$\|\Pi_{\text{win}} z_j^{(1)} - \Pi_{\text{win}}\widetilde{z}_j\|_\infty \le CK \exp\left(-\frac{c\chi}{K^2}\right).$$

By construction, $\Pi_{\text{win}}\widetilde{z}_j = \Gamma_K(\Sigma_j X)$, and therefore

$$\boxed{\|\Pi_{\text{win}} z_j^{(1)} - \Gamma_K(\Sigma_j X)\|_\infty \le CK \exp\left(-\frac{c\chi}{K^2}\right).} \tag{B.2}$$

It remains to approximate the map

$$(e_{X_{j-K}}, \dots, e_{X_{j+K}}) \longmapsto f_K(\Sigma_j X)$$

for which it suffices to approximate $(e_{X_{-K}}, \dots, e_{X_K}) \longmapsto f_K(X)$ in sup norm. For every index $\gamma \in \mathfrak{A}_{K,r}$,

$$b_\gamma(X) = \prod_{\ell \in \operatorname{supp}\gamma} \psi_{\gamma_\ell}(X_\ell), \qquad \psi_a(x) = v_a^\top e_x$$

for some v_a. Thus each factor $\psi_a(X_\ell)$ is a linear map from the corresponding one-hot vector, while b_γ is a product of at most r such bounded scalars.

For each $\gamma \in \mathfrak{A}_{K,r}$, write $k_\gamma := |\operatorname{supp}\gamma| = \operatorname{ord}(\gamma) \le r$. The product-network construction in the proof of

By Proposition 3 of Yarotsky (2017) (see also the k_γ-fold product construction in the proof of Theorem 8 of Nishimura & Suzuki (2024)), for every $0 < \delta < 1$ there exists a ReLU FNN $\mathrm{Prod}_{k_\gamma,\delta} : [-1,1]^{k_\gamma} \to \mathbb{R}$ such that

$$\sup_{z\in[-1,1]^{k_\gamma}} \left|\mathrm{Prod}_{k_\gamma,\delta}(z) - \prod_{i=1}^{k_\gamma} z_i\right| \le \delta.$$

Since $k_\gamma \le r$ and r is fixed, it may be chosen with width $O_r(1)$, depth and sparsity $O_r(\log(1/\delta))$, and weights bounded uniformly in δ.

Applying this network to the exactly computed linear factors $\psi_{\gamma_\ell}(X_\ell) = v_{\gamma_\ell}^\top e_{X_\ell}$ gives an FNN $\widehat{b}_{\gamma,\delta}$ satisfying

$$\|\widehat{b}_{\gamma,\delta} - b_\gamma\|_\infty \le \delta$$

with $O_r(\log(1/\delta))$ nonzero parameters.

The resulting position-wise FNN is

$$\widehat{f}_K := \sum_{\gamma \in \mathfrak{A}_{K,r}} d_\gamma \widehat{b}_{\gamma,\delta_K}.$$

where δ_K is to be chosen later.

Bounding the approximation error of the FNN. The retained coefficient mass is uniformly bounded, i.e., $\sum_{\gamma \in \mathfrak{A}_{K,r}} |d_\gamma| \le C'$. Set $\varepsilon_K := (K+1)^\kappa 2^{-\beta K}$ and $\delta_K := \varepsilon_K / C'$. Then we get the error bound

$$\|f_K - \widehat{f}_K\|_\infty \le \varepsilon_K. \tag{B.3}$$

Since $\log(1/\delta_K) = O(K)$ and $G_r(K) = O(K^r)$, we get the network size bounds

$$L_{\mathrm{net},K} \le C(K+1), \qquad W_{\mathrm{net},K} \le C(K+1)^r, \qquad S_{\mathrm{net},K} \le C(K+1)^{r+1}.$$

Combining the bounds B.1 and B.3,

$$\boxed{\|f - \widehat{f}_K\|_\infty \le C\varepsilon_K}$$

Using the lipschitzness of $\hat{f}_K$. Let

$$\Lambda_K := \mathrm{Lip}(\widehat{f}_K).$$

Choose the parameter

$$\chi = \frac{K^2}{c} \log\left(\frac{CK\Lambda_K}{\varepsilon_K}\right).$$

Then the error in B.2, propagated through the FNN, satisfies

$$\left|\widehat{f}_K(\Pi_{\mathrm{win}} z_j^{(1)}) - \widehat{f}_K(\Pi_{\mathrm{win}} \widetilde{z}_j)\right| \le \varepsilon_K.$$

Combine the approximations: Define the Transformer by

$$\widehat{F}_{K,j}(X) := \widehat{f}_K(\Pi_{\mathrm{win}} z_j^{(1)}).$$

Since

$$\Pi_{\mathrm{win}} \widetilde{z}_j = \Gamma_K(\Sigma_j X),$$

we have

$$\begin{aligned} |\widehat{F}_{K,j}(X) - F_j^\circ(X)| &\le \left|\widehat{f}_K(\Pi_{\mathrm{win}} z_j^{(1)}) - \widehat{f}_K(\Pi_{\mathrm{win}} \widetilde{z}_j)\right| \\ &\quad + \left|\widehat{f}_K(\Gamma_K(\Sigma_j X)) - f(\Sigma_j X)\right| \\ &\le C\varepsilon_K. \end{aligned}$$

Therefore

$$\sup_{j \in \mathbb{Z}} \|\widehat{F}_{K,j} - F_j^\circ\|_\infty \le C(K+1)^\kappa 2^{-\beta K}.$$

The depth, width, sparsity, parameter-magnitude, and parameter-count bounds follow from the same attention and product-network constructions.

B.1 Proof of Consequences

Proof of Corollary A.2. Write

$$a_{\mathrm{par}} := \max\{r+1, 3\}, \qquad c_\beta := \beta \log 2.$$

By Theorem 4.1, there are constants $C_{\mathrm{app}}, C_{\mathrm{par}} > 0$, depending only on q, ν, r, β, τ, such that

$$\sup_{j \in \mathbb{Z}} \|\widehat{F}_{K,j} - F_j^\circ\|_\infty \le C_{\mathrm{app}}(K+1)^\kappa e^{-c_\beta K}, \qquad N_K \le C_{\mathrm{par}}\big\{(K+1)^{r+1} + q^2(K+1)^3\big\}. \tag{B.4}$$

Choose K using parameter count Since $a_{\mathrm{par}} = \max\{r+1, 3\}$,

$$N_K \le C_N (K+1)^{a_{\mathrm{par}}}, \qquad C_N := C_{\mathrm{par}}(1+q^2).$$

For all sufficiently large N, choose

$$K_N := \left\lfloor \left(\frac{N}{C_N} \right)^{1/a_{\mathrm{par}}} \right\rfloor - 1.$$

Then $K_N \ge 2$ and $N_{K_N} \le N$. Moreover, after increasing N if necessary,

$$\frac{1}{2} C_N^{-1/a_{\mathrm{par}}} N^{1/a_{\mathrm{par}}} \le K_N \le K_N + 1 \le C_N^{-1/a_{\mathrm{par}}} N^{1/a_{\mathrm{par}}}.$$

Substitute into the approximation bound Using (B.4),

$$\begin{aligned} \sup_{j \in \mathbb{Z}} \|\widehat{F}_{N,j} - F_j^\circ\|_\infty &\le C_{\mathrm{app}} (K_N + 1)^\kappa e^{-c_\beta K_N} \\ &\le C N^{\kappa/a_{\mathrm{par}}} \exp\Big(-c N^{1/a_{\mathrm{par}}} \Big), \end{aligned}$$

where one may take, for example,

$$c = \frac{\beta \log 2}{2 C_N^{1/a_{\mathrm{par}}}},$$

and $C > 0$ depends only on q, ν, r, β, τ. Since μ is a probability measure, $\|h\|_{L^2(\mu)} \le \|h\|_\infty$, so the same bound holds in $L^2(\mu)$. □

Proof of Corollary A.3. Let

$$A_\varepsilon := \log_2 \frac{1}{\varepsilon}.$$

By Theorem 4.1, it suffices to choose K so that

$$C_{\mathrm{app}} (K+1)^\kappa 2^{-\beta K} \le \varepsilon, \tag{B.5}$$

where C_{app} depends only on q, ν, r, β, τ.

Choose K_ε For sufficiently small ε, take

$$K_\varepsilon = \left\lceil \frac{1}{\beta} \left(A_\varepsilon + \kappa \log_2 A_\varepsilon + C_* \right) \right\rceil$$

where $C_* > 0$ is a sufficiently large constant depending only on q, ν, r, β, τ. Since $K_\varepsilon + 1 = O(A_\varepsilon)$,

$$\begin{aligned} C_{\mathrm{app}} (K_\varepsilon + 1)^\kappa 2^{-\beta K_\varepsilon} &\le C A_\varepsilon^\kappa 2^{-A_\varepsilon} A_\varepsilon^{-\kappa} 2^{-C_*} \\ &\le \varepsilon \end{aligned}$$

after choosing C_* large enough. Hence

$$K_\varepsilon = \frac{1}{\beta} \log_2 \frac{1}{\varepsilon} + O\left(\log \log \frac{1}{\varepsilon} \right)$$

where the constant hidden in $O(\cdot)$ depends only on q, ν, r, β, τ.

Convert K_ε into parameter complexity Let $a_{\mathrm{par}} := \max\{r+1, 3\}$. From the parameter bound in Theorem 4.1,

$$N_{K_\varepsilon} \le C_{\mathrm{par}} \big\{ (K_\varepsilon + 1)^{r+1} + q^2 (K_\varepsilon + 1)^3 \big\} \le C (K_\varepsilon + 1)^{a_{\mathrm{par}}}$$

Since $K_\varepsilon = O(\log(1/\varepsilon))$,

$$N(\varepsilon) \le C \left(\log \frac{1}{\varepsilon} \right)^{\max\{r+1, 3\}}$$

where $C > 0$ depends only on q, ν, r, β, τ. □

C TOOLS FROM CTMC THEORY

C.1 LIKELIHOOD RATIOS OF PATH MEASURES FOR FINITE-STATE CTMCS

Let $(Y_s)_{0\le s\le T}$ be a finite-state, time-inhomogeneous CTMC on a finite state space $\mathcal{S}$. Consider two path laws $\mathbb{P}^r$ and $\mathbb{P}^q$ with the same initial law μ and off-diagonal jump rates

$$r_s(a,b), \qquad q_s(a,b), \qquad a,b\in\mathcal{S}, \quad a\neq b,$$

and define their exit rates by

$$\lambda_s^r(a) := \sum_{b\neq a} r_s(a,b), \qquad \lambda_s^q(a) := \sum_{b\neq a} q_s(a,b).$$

Assume that $r_s(a,b)>0$ implies $q_s(a,b)>0$.

Consider a path ω with states $a_0,a_1,\dots,a_N$ and jump times $0<\tau_1<\cdots<\tau_N<T$, and put $\tau_0=0$, $\tau_{N+1}=T$. Since $Y_s=a_k$ for $s\in[\tau_k,\tau_{k+1})$, the path density under $\mathbb{P}^r$ is

$$d\mathbb{P}^r(\omega) = \mu(a_0)\prod_{k=0}^{N}\exp\left\{-\int_{\tau_k}^{\tau_{k+1}}\lambda_s^r(a_k)\,ds\right\}\prod_{k=1}^{N} r_{\tau_k}(a_{k-1},a_k), \tag{C.1}$$

and similarly

$$d\mathbb{P}^q(\omega) = \mu(a_0)\prod_{k=0}^{N}\exp\left\{-\int_{\tau_k}^{\tau_{k+1}}\lambda_s^q(a_k)\,ds\right\}\prod_{k=1}^{N} q_{\tau_k}(a_{k-1},a_k). \tag{C.2}$$

Dividing (C.1) by (C.2) gives

$$\begin{aligned}\frac{d\mathbb{P}^r}{d\mathbb{P}^q}(\omega) &= \exp\left\{-\sum_{k=0}^{N}\int_{\tau_k}^{\tau_{k+1}}\left[\lambda_s^r(a_k)-\lambda_s^q(a_k)\right]ds\right\}\prod_{k=1}^{N}\frac{r_{\tau_k}(a_{k-1},a_k)}{q_{\tau_k}(a_{k-1},a_k)}\\ &= \exp\left\{-\int_0^T\left[\lambda_s^r(Y_s)-\lambda_s^q(Y_s)\right]ds\right\}\prod_{k=1}^{N}\frac{r_{\tau_k}(a_{k-1},a_k)}{q_{\tau_k}(a_{k-1},a_k)}.\end{aligned} \tag{C.3}$$

Hence

$$\log\frac{d\mathbb{P}^r}{d\mathbb{P}^q}(\omega) = -\int_0^T\left[\lambda_s^r(Y_s)-\lambda_s^q(Y_s)\right]ds + \sum_{k=1}^{N}\log\frac{r_{\tau_k}(a_{k-1},a_k)}{q_{\tau_k}(a_{k-1},a_k)}. \tag{C.4}$$

For the reverse masking processes below, one may apply (C.4) to the increasing-time parameterization $Y_s=Z_{1-s}$ and then change variables back to the forward-time parameter $t=1-s$.

C.2 REVERSE RATES FOR THE MASKING PROCESS

For $z\in\mathcal{E}^L$, $i\in\mathcal{M}(z)$, and $a\in\mathcal{A}$, write

$$z^{i\leftarrow a} := (z_1,\dots,z_{i-1},a,z_{i+1},\dots,z_L).$$

Fix a clean word $x\in\mathcal{A}^L$. Let $\mathbb{Q}^x$ denote the law of the true reverse masking process conditional on $X=x$. For a fixed coordinate i, let ξ_i be its masking time, so that $\mathbb{P}(\xi_i>t)=\alpha_t$. Conditional on $Z_t=z$ and $X=x$, coordinate $i\in\mathcal{M}(z)$ is already masked, and therefore

$$\begin{aligned}&\mathbb{P}\left(Z_{t-h}=z^{i\leftarrow a}\mid Z_t=z,\ X=x\right)\\ &\quad=\mathbf{1}_{\{a=x_i\}}\mathbb{P}(t-h<\xi_i\le t\mid \xi_i\le t) = \mathbf{1}_{\{a=x_i\}}\frac{\alpha_{t-h}-\alpha_t}{1-\alpha_t}.\end{aligned} \tag{C.5}$$

Consequently,

$$r_t^x\left(z,z^{i\leftarrow a}\right) := \lim_{h\downarrow 0}\frac{1}{h}\mathbb{P}\left(Z_{t-h}=z^{i\leftarrow a}\mid Z_t=z,\ X=x\right) = w(t)\mathbf{1}_{\{a=x_i\}}. \tag{C.6}$$

Now average over the clean sequence conditional on $Z_t = z$. Since the masking pattern is independent of the token values,

$$\mathbb{P}(X_i = a \mid Z_t = z) = \pi_i(a \mid z).$$

Thus the unconditional true reverse process has rates

$$r_t\left(z, z^{i \leftarrow a}\right) = w(t)\pi_i(a \mid z). \tag{C.7}$$

Given a strictly positive predictor p, let $\mathbb{P}_p$ denote the learned reverse CTMC whose rates are

$$r_t^p\left(z, z^{i \leftarrow a}\right) = w(t)p_i(a \mid z, t). \tag{C.8}$$

Since $\pi_i(\cdot \mid z)$ and $p_i(\cdot \mid z, t)$ are probability distributions, the three corresponding exit rates are

$$\lambda_t^x(z) = \sum_{i \in \mathcal{M}(z)} \sum_{a \in \mathcal{A}} w(t)\mathbf{1}_{\{a = x_i\}} = w(t)|\mathcal{M}(z)|, \tag{C.9}$$

$$\lambda_t(z) = \sum_{i \in \mathcal{M}(z)} \sum_{a \in \mathcal{A}} w(t)\pi_i(a \mid z) = w(t)|\mathcal{M}(z)|, \tag{C.10}$$

$$\lambda_t^p(z) = \sum_{i \in \mathcal{M}(z)} \sum_{a \in \mathcal{A}} w(t)p_i(a \mid z, t) = w(t)|\mathcal{M}(z)|. \tag{C.11}$$

Thus the true and learned reverse processes have the same total unmasking rate; they differ only in the distribution of the token inserted at an unmasking event.

C.3 Excess risk of an MDLM

The following identity and its proof have several precursors in literature. A version of this for the discrete score has been stated frequently, see for instance Campbell et al. (2022), Lou et al. (2024), Ren et al. (2025a) and Jeon et al. (2025). Nevertheless, we provide the proof below for completeness. The identity follows by evaluating the KL divergence between paths in two different ways.

Theorem C.1 (MDLM excess-risk)**.** *For every strictly positive predictor p with $R(p) < \infty$,*

$$R(p) - H(P) = \int_0^1 w(t)\, \mathbb{E}\left[\sum_{i \in \mathcal{M}(Z_t)} D_{\mathrm{KL}}\left(\pi_i(\cdot \mid Z_t) \| p_i(\cdot \mid Z_t, t)\right)\right] dt. \tag{C.12}$$

Proof. Let

$$\mathbb{J}_\Omega(d\omega) := \sum_{x \in \mathcal{A}^L} P(x)\mathbb{Q}^x(d\omega)$$

be the unconditional true reverse path law, and let $\Phi(\omega) = Z_0$ denote the clean endpoint of a reverse path.

First evaluation of the path-space KL. Fix $x \in \mathrm{supp}(P)$. By (C.6), (C.8), and (C.11), the exit-rate term in (C.4) vanishes. At a jump that unmasks coordinate i, the true process necessarily inserts x_i, and hence the jump-rate ratio is

$$\frac{w(t)}{w(t)p_i(x_i \mid z, t)} = \frac{1}{p_i(x_i \mid z, t)}.$$

Therefore

$$\log \frac{d\mathbb{Q}^x}{d\mathbb{P}_p}(\omega) = \sum_{\tau \in J(\omega)} -\log p_{i(\tau)}\big(x_{i(\tau)} \mid Z_\tau, \tau\big), \tag{C.13}$$

where $J(\omega)$ is the set of jump times and $i(\tau)$ is the coordinate unmasked at time τ. Taking expectation under $\mathbb{Q}^x$ and using the jump intensity $w(t)$ at every masked coordinate gives

$$D_{\mathrm{KL}}(\mathbb{Q}^x \| \mathbb{P}_p) = \int_0^1 w(t) \sum_{z \in \mathcal{E}^L} q_t(z \mid x) \sum_{i \in \mathcal{M}(z)} -\log p_i(x_i \mid z, t)\, dt. \tag{C.14}$$

Under $\mathbb{Q}^x$, the clean endpoint satisfies $\Phi(\omega) = x$ almost surely. Hence, for $\mathbb{J}_\Omega$-almost every ω,

$$\frac{d\mathbb{J}_\Omega}{d\mathbb{P}_p}(\omega) = P(\Phi(\omega)) \frac{d\mathbb{Q}^{\Phi(\omega)}}{d\mathbb{P}_p}(\omega). \tag{C.15}$$

Taking logarithms and expectation with respect to $\mathbb{J}_\Omega$, and then using (C.14), yields

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{J}_\Omega \| \mathbb{P}_p) &= \sum_{x \in \mathcal{A}^L} P(x) \log P(x) + \sum_{x \in \mathcal{A}^L} P(x) D_{\mathrm{KL}}(\mathbb{Q}^x \| \mathbb{P}_p) \\ &= -H(P) + R(p). \end{aligned} \tag{C.16}$$

Second evaluation of the same path-space KL. The reverse jump rates of $\mathbb{J}_\Omega$ and $\mathbb{P}_p$ are, by (C.7) and (C.8),

$$w(t)\pi_i(a \mid z) \qquad \text{and} \qquad w(t)p_i(a \mid z, t),$$

respectively. Their exit rates agree by (C.11), so (C.4) gives

$$\log \frac{d\mathbb{J}_\Omega}{d\mathbb{P}_p}(\omega) = \sum_{\tau \in J(\omega)} \log \frac{\pi_{i(\tau)}(a(\tau) \mid Z_\tau)}{p_{i(\tau)}(a(\tau) \mid Z_\tau, \tau)}, \tag{C.17}$$

where $a(\tau)$ is the token inserted at time τ. Conditional on $Z_t = z$, a jump inserting a at coordinate $i \in \mathcal{M}(z)$ has true rate $w(t)\pi_i(a \mid z)$. Taking expectation in (C.17) therefore gives

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{J}_\Omega \| \mathbb{P}_p) &= \int_0^1 w(t)\, \mathbb{E}\left[\sum_{i \in \mathcal{M}(Z_t)} \sum_{a \in \mathcal{A}} \pi_i(a \mid Z_t) \log \frac{\pi_i(a \mid Z_t)}{p_i(a \mid Z_t, t)} \right] dt \\ &= \int_0^1 w(t)\, \mathbb{E}\left[\sum_{i \in \mathcal{M}(Z_t)} D_{\mathrm{KL}}\left(\pi_i(\cdot \mid Z_t) \| p_i(\cdot \mid Z_t, t)\right) \right] dt. \end{aligned} \tag{C.18}$$

Equating (C.16) and (C.18) proves (C.12). □

D Proof of Theorem 6.1 and Corollary A.4

Note that this theorem is a standard consequence of the fact that $R(p)$ is a Negative ELBO (see Appendix B.2 of Sahoo et al. (2024)), and has occurred in other forms in literature. We provide a proof using CTMC theory.

Proof of Theorem 6.1. Apply the data-processing inequality to the measurable endpoint map Φ. Since

$$\Phi_\# \mathbb{J}_\Omega = P, \qquad \Phi_\# \mathbb{P}_p^{(\infty)} = P_p^{(\infty)},$$

we obtain

$$D_{\mathrm{KL}}\left(P \middle\| P_p^{(\infty)}\right) \le D_{\mathrm{KL}}\left(\mathbb{J}_\Omega \middle\| \mathbb{P}_p^{(\infty)}\right) = R(p) - H(P).$$

where we have used Eqn C.16 in the equality.

For the finite-step sampler, data processing similarly gives

$$D_{\mathrm{KL}}\left(P \middle\| P_p^{(T)}\right) \le D_{\mathrm{KL}}\left(\mathbb{J}_\Omega \middle\| \mathbb{P}_p^{(T)}\right).$$

By the definition of $\mathrm{Disc}_p(T)$,

$$D_{\mathrm{KL}}\left(\mathbb{J}_\Omega \middle\| \mathbb{P}_p^{(T)}\right) \le D_{\mathrm{KL}}\left(\mathbb{J}_\Omega \middle\| \mathbb{P}_p^{(\infty)}\right) + \mathrm{Disc}_p(T),$$

and hence

$$D_{\mathrm{KL}}\left(P \middle\| P_p^{(T)}\right) \le R(p) - H(P) + \mathrm{Disc}_p(T).$$

□

Proof of Corollary A.4. By definition,

$$\begin{aligned} R(\widehat{p}_{K,m}) - H(P) &= \left[\inf_{p \in \mathcal{P}_K} R(p) - H(P) \right] + \left[R(\widehat{p}_{K,m}) - \inf_{p \in \mathcal{P}_K} R(p) \right] \\ &= \mathrm{Approx}(\mathcal{P}_K) + \mathrm{Est}_{K,m}\,. \end{aligned}$$

Substituting this identity into the two inequalities of Theorem 6.1 proves the result. □

E Proof of Theorem 6.2

Fix $K \geq 2$. Suppose first that $R_\lambda > 0$.

Approximate the smoothed posterior logits For each $a \in \mathcal{A}$, define

$$g_a^\lambda := \frac{f_a^\lambda}{R_\lambda}.$$

By (5.7), $\|g_a^\lambda\|_{\mathcal{B}_{\beta,\tau}^{(r)}} \leq 1$. Hence Theorem 4.1, applied to the shifted targets $g_a^\lambda \circ \Sigma_i$, gives a one-block Transformer $\widehat{G}_{a,K}$ such that

$$\sup_{i \in \mathbb{Z}} \|\widehat{G}_{a,K,i} - g_a^\lambda \circ \Sigma_i\|_\infty \leq C_{\mathrm{tr}}(K+1)^\kappa 2^{-\beta K}.$$

Set $\widehat{F}_{a,K} := R_\lambda \widehat{G}_{a,K}$. Using $\ell_{i,a}^\lambda(z) = f_a^\lambda(\Sigma_i \iota z)$,

$$\max_{\substack{z \in \mathcal{Z}_P,\ i \in \mathcal{M}(z) \\ a \in \mathcal{A}}} \left| \widehat{F}_{a,K,i}(\iota z) - \ell_{i,a}^\lambda(z) \right| \leq \rho_K, \qquad \rho_K := \rho_{\lambda,\mathrm{Tr}}(K). \tag{E.1}$$

Convert logit error into a conditional KL bound Fix $z \in \mathcal{Z}_P$ and $i \in \mathcal{M}(z)$, and write

$$q_a := \pi_i^\lambda(a \mid z), \qquad v_a := \widehat{F}_{a,K,i}(\iota z), \qquad Z_K := \sum_{b \in \mathcal{A}} e^{v_b}.$$

Since $\ell_{i,a}^\lambda(z) = \log q_a$, (E.1) implies $|v_a - \log q_a| \leq \rho_K$. Therefore

$$e^{-\rho_K} q_a \leq e^{v_a} \leq e^{\rho_K} q_a, \qquad e^{-\rho_K} \leq Z_K \leq e^{\rho_K}.$$

Define

$$p_{K,i}(a \mid z, t) = p_{K,i}(a \mid z) := \frac{e^{v_a}}{Z_K}.$$

Then

$$\left| \log \frac{q_a}{p_{K,i}(a \mid z)} \right| = |(\log q_a - v_a) + \log Z_K| \leq 2\rho_K. \tag{E.2}$$

Moreover,

$$q_a = (1-\lambda)\pi_i(a \mid z) + \frac{\lambda}{M} \geq (1-\lambda)\pi_i(a \mid z),$$

so

$$D_{\mathrm{KL}}\big(\pi_i(\cdot \mid z) \big\| \pi_i^\lambda(\cdot \mid z)\big) \leq \log \frac{1}{1-\lambda}.$$

Combining this with (E.2),

$$\begin{aligned} D_{\mathrm{KL}}(\pi_i(\cdot \mid z) \| p_{K,i}(\cdot \mid z)) &= D_{\mathrm{KL}}\big(\pi_i(\cdot \mid z) \big\| \pi_i^\lambda(\cdot \mid z)\big) + \sum_{a \in \mathcal{A}} \pi_i(a \mid z) \log \frac{q_a}{p_{K,i}(a \mid z)} \\ &\leq \log \frac{1}{1-\lambda} + 2\rho_K. \end{aligned} \tag{E.3}$$

Verify finite risk By (5.6) and (E.1),

$$-A_\lambda - \rho_K \leq v_a \leq \rho_K,$$

hence

$$p_{K,i}(a \mid z) \geq \frac{e^{-A_\lambda - 2\rho_K}}{M}.$$

Thus $-\log p_{K,i}(a \mid z)$ is uniformly bounded, and moreover,

$$\int_0^1 w(t)\, \mathbb{E}|\mathcal{M}(Z_t)|\, dt = \int_0^1 w(t) L(1-\alpha_t) = L < \infty \tag{E.4}$$

This gives

$$\begin{aligned} R(p_K) = \mathbb{E}_{X \sim P} \int_0^1 w(t) \\ \times \sum_{z \in \mathcal{E}^L} q_t(z \mid X) \sum_{i \in \mathcal{M}(z)} -\log p_{K,i}(X_i \mid z, t)\, dt < \infty \end{aligned} \tag{E.5}$$

Apply the excess-risk identity By Theorem C.1 and (E.3),

$$R(p_K) - H(P) = \int_0^1 w(t)\,\mathbb{E}\left[\sum_{i\in\mathcal{M}(Z_t)} D_{\mathrm{KL}}(\pi_i(\cdot \mid Z_t)\|p_{K,i}(\cdot \mid Z_t))\right] dt$$
$$\le \left(\log\frac{1}{1-\lambda} + 2\rho_K\right)\int_0^1 w(t)\mathbb{E}|\mathcal{M}(Z_t)|\,dt$$
$$= L\log\frac{1}{1-\lambda} + 2L\rho_K. \tag{E.6}$$

Since $\rho_K = C_{\mathrm{tr}}R_\lambda(K+1)^\kappa 2^{-\beta K}$,

$$\mathrm{Approx}(\mathcal{P}^{\mathrm{bd}}_{K,\lambda}) \le R(p_K) - H(P) \le L\log\frac{1}{1-\lambda} + 2LC_{\mathrm{tr}}R_\lambda(K+1)^\kappa 2^{-\beta K}.$$

Read off the architecture bounds Theorem 4.1 gives, for each scalar branch,

$$U_K = K, \qquad H_K = 2K+1, \qquad D_K = Q + Q(2K+1) + 2,$$

together with

$$L_{\mathrm{net},K} \le C_{\mathrm{tr}}(K+1), \qquad W_{\mathrm{net},K} \le C_{\mathrm{tr}}(K+1)^r, \qquad S_{\mathrm{net},K} \le C_{\mathrm{tr}}(K+1)^{r+1}.$$

Multiplying the final affine map by R_λ does not change the depth, width, sparsity, or number of parameters, and yields

$$B_{\mathrm{net},K} \le C_{\mathrm{tr}}(1\vee R_\lambda)(K+1)^3\log(K+2).$$

Summing the parameter counts of the M branches gives

$$N_{K,M} \le C_{\mathrm{tr}}M\left((K+1)^{r+1} + Q^2(K+1)^3\right).$$

If $R_\lambda = 0$, then $f^\lambda_a = 0$ for every $a \in \mathcal{A}$. The zero-output Transformers represent these logits exactly, so the same argument applies with $\rho_{\lambda,\mathrm{Tr}}(K) = 0$.

F Proof of Theorem 6.3 and Its Consequences

For any $p = p_{\mathbf{F}} \in \mathcal{P}^{\mathrm{bd}}_{K,\lambda}$, the logit bounds

$$-A_\lambda - \rho_K \le F_{a,i}(\iota z) \le \rho_K$$

imply, for every $z \in \mathcal{Z}_P$, $i \in \mathcal{M}(z)$, and $a \in \mathcal{A}$,

$$p_i(a \mid z) = \frac{e^{F_{a,i}(\iota z)}}{\sum_{b\in\mathcal{A}} e^{F_{b,i}(\iota z)}} \ge \frac{e^{-A_\lambda-\rho_K}}{Me^{\rho_K}} = \frac{e^{-A_\lambda-2\rho_K}}{M}.$$

Hence

$$0 \le -\log p_i(a \mid z) \le A_\lambda + 2\rho_K + \log M.$$

Therefore, for every $x \in \mathrm{supp}(P)$,

$$0 \le J_p(x) \le \big(A_\lambda + 2\rho_K + \log M\big)\int_0^1 w(t)\sum_{z\in\mathcal{E}^L} q_t(z \mid x)\,|\mathcal{M}(z)|\,dt$$
$$= L\big(A_\lambda + 2\rho_K + \log M\big)\int_0^1 w(t)(1-\alpha_t)\,dt$$
$$= -L\big(A_\lambda + 2\rho_K + \log M\big)\int_0^1 \alpha_t'\,dt$$
$$= L\big(A_\lambda + 2\rho_K + \log M\big) = \mathcal{B}_{K,\lambda},$$

where we used

$$\sum_{z\in\mathcal{E}^L} q_t(z \mid x)|\mathcal{M}(z)| = L(1-\alpha_t), \qquad w(t) = -\frac{\alpha_t'}{1-\alpha_t},$$

and $\alpha(0) = 1$, $\alpha(1) = 0$. The first equality is just the linearity of expectation applied to $\mathcal{E}_{z\sim q_t(\cdot|x)}|\mathcal{M}(z)|$.

Hence, for every $p \in \mathcal{P}^{\mathrm{bd}}_{K,\lambda}$, the per-example loss satisfies $0 \le J_p(x) \le B := \mathcal{B}_{K,\lambda}$. Then equation (5.8) in Theorem 5.7 of Sutherland (2026) gives, with probability at least $1-\delta$,

$$R(\widehat{p}_m) - \inf_{p\in\mathcal{P}^{\mathrm{bd}}_{K,\lambda}} R(p) \le \mathbb{E} \sup_{p\in\mathcal{P}^{\mathrm{bd}}_{K,\lambda}} \{R(p) - \widehat{R}_m(p)\} + B\sqrt{\frac{2}{m}\log\frac{2}{\delta}}.$$

Centering the loss class at $B/2$ does not change either the empirical process or its covering numbers, and gives a class of L^∞-radius $B/2$. Proposition 2.6 of Chewi et al. (2025) therefore yields

$$\mathbb{E}\sup_p\{R(p) - \widehat{R}_m(p)\} \le C_{\mathrm{ch}} \inf_{0<\varepsilon_0\le B/2}\left\{\varepsilon_0 + \frac{1}{\sqrt{m}}\int_{\varepsilon_0}^{B/2}\sqrt{\log N(\varepsilon, \mathscr{L}_{K,\lambda}, \|\cdot\|_\infty)}\,d\varepsilon\right\}.$$

It remains to bound the covering number of the loss class.

Loss-to-model lipschitzness. For $y \in \mathcal{A}$, define the cross-entropy as a function of the logits,

$$\phi_y(v) := -v_y + \log\sum_{a\in\mathcal{A}} e^{v_a}, \qquad v \in \mathbb{R}^M.$$

Since

$$\nabla\phi_y(v) = \mathrm{Softmax}(v) - e_y,$$

we have

$$\|\nabla\phi_y(v)\|_1 = \sum_{a\ne y}\mathrm{Softmax}(v)_a + 1 - \mathrm{Softmax}(v)_y = 2\big(1 - \mathrm{Softmax}(v)_y\big) \le 2.$$

Hence, by the mean-value theorem and duality of ℓ^1 and ℓ^∞,

$$|\phi_y(v) - \phi_y(v')| \le 2\|v - v'\|_\infty. \tag{F.1}$$

Now fix $x \in \mathrm{supp}(P)$, $z \in \mathcal{Z}_P$, and $i \in \mathcal{M}(z)$. Since

$$-\log p_{\mathbf{F},i}(x_i \mid z) = \phi_{x_i}\Big(\big(F_{a,i}(\iota z)\big)_{a\in\mathcal{A}}\Big),$$

(F.1) gives

$$|-\log p_{\mathbf{F},i}(x_i \mid z) + \log p_{\mathbf{G},i}(x_i \mid z)| \le 2\, d_{\mathrm{logit}}(\mathbf{F},\mathbf{G}).$$

Therefore, using (A.6),

$$\begin{aligned}
|J_{p_{\mathbf{F}}}(x) - J_{p_{\mathbf{G}}}(x)| &\le 2\, d_{\mathrm{logit}}(\mathbf{F},\mathbf{G}) \int_0^1 w(t) \sum_{z\in\mathcal{E}^L} q_t(z \mid x)|\mathcal{M}(z)|\,dt \\
&= 2L\, d_{\mathrm{logit}}(\mathbf{F},\mathbf{G}) \int_0^1 w(t)(1-\alpha_t)\,dt \\
&= 2L\, d_{\mathrm{logit}}(\mathbf{F},\mathbf{G})
\end{aligned}$$

Taking the supremum over $x \in \mathrm{supp}(P)$ yields

$$\|J_{p_{\mathbf{F}}} - J_{p_{\mathbf{G}}}\|_\infty \le 2L\, d_{\mathrm{logit}}(\mathbf{F},\mathbf{G}) \tag{F.2}$$

Consequently, using an M-fold product cover of the scalar Transformer class,

Covering Number of the Loss using Covering Number of the Model Recall that

$$\mathscr{L}_{K,\lambda} = \left\{ x \mapsto \int_0^1 w(t) \sum_{z \in \mathcal{E}^L} q_t(z \mid x) \sum_{i \in \mathcal{M}(z)} -\log p_{\mathbf{F},i}(x_i \mid z)\, dt \;:\; \begin{array}{l} \mathbf{F} = (F_a)_{a\in\mathcal{A}} \in \mathcal{T}_K^M, \\ p_{\mathbf{F}} = \mathrm{Softmax}(\mathbf{F}), \\ -A_\lambda - \rho_K \le F_{a,i}(\iota z) \le \rho_K \\ \text{for all } z \in \mathcal{Z}_P,\ i \in \mathcal{M}(z),\ a \in \mathcal{A} \end{array} \right\}.$$

By the inequality F.2, an $\varepsilon/(2L)$-cover of the vectors $(F_a)_{a\in\mathcal{A}}$ appearing in the definition above, with respect to d_{logit}, induces an ε-cover of $\mathscr{L}_{K,\lambda}$. Hence

$$\begin{aligned} N(\varepsilon, \mathscr{L}_{K,\lambda}, \|\cdot\|_\infty) &\le N\Big(\frac{\varepsilon}{2L}, \Big\{ \mathbf{F} = (F_a)_{a\in\mathcal{A}} \in \mathcal{T}_K^M : \begin{array}{l} -A_\lambda - \rho_K \le F_{a,i}(\iota z) \le \rho_K \\ \text{for all } z \in \mathcal{Z}_P,\ i \in \mathcal{M}(z),\ a \in \mathcal{A} \end{array} \Big\}, d_{\mathrm{logit}}\Big) \\ &\le N\Big(\frac{\varepsilon}{4L}, \mathcal{T}_K^M, d_{\mathrm{logit}}\Big). \end{aligned}$$

Here the second inequality uses the general fact that, for $A \subseteq B$,

$$N(\eta, A, d) \le N(\eta/2, B, d).$$

Finally, if $\{T_1, \dots, T_N\}$ is an $\varepsilon/(4L)$-cover of $\mathcal{T}_K$ in $\|\cdot\|_{\infty,\mathrm{TS}}$, then its M-fold product is an $\varepsilon/(4L)$-cover of $\mathcal{T}_K^M$ in d_{logit}, since (A.5) gives

$$d_{\mathrm{logit}}(\mathbf{F}, \mathbf{G}) \le \max_{a\in\mathcal{A}} \|F_a - G_a\|_{\infty,\mathrm{TS}}.$$

Therefore

$$N\Big(\frac{\varepsilon}{4L}, \mathcal{T}_K^M, d_{\mathrm{logit}}\Big) \le N\Big(\frac{\varepsilon}{4L}, \mathcal{T}_K, \|\cdot\|_{\infty,\mathrm{TS}}\Big)^M$$

and hence

$$N(\varepsilon, \mathscr{L}_{K,\lambda}, \|\cdot\|_\infty) \le N\Big(\frac{\varepsilon}{4L}, \mathcal{T}_K, \|\cdot\|_{\infty,\mathrm{TS}}\Big)^M$$

Applying the Takakura & Suzuki (2023) formula for covering number By Takakura & Suzuki (2023, Theorem 5.3), written in the notation of Section 3,

$$\log N(\eta, \mathcal{T}(N_{\mathrm{blk}}, U, D, H, L_{\mathrm{net}}, W_{\mathrm{net}}, S_{\mathrm{net}}, B_{\mathrm{cov}}; \mathrm{PE}), \|\cdot\|_\infty) \lesssim N_{\mathrm{blk}}^3 L_{\mathrm{net}} \left(S_{\mathrm{net}} + HD^2\right) \log\left(\frac{DHL_{\mathrm{net}}W_{\mathrm{net}}B_{\mathrm{cov}}}{\eta}\right)$$

Thus, choosing a universal constant C_{TS} for the implicit constant, using $N_{\mathrm{blk}} = 1$, $B_{\mathrm{cov},K} \ge 1$, and $\|\mathrm{PE}_K\|_\infty \le B_{\mathrm{cov},K}$, and recalling

$$V_K = C_{\mathrm{TS}} M L_{\mathrm{net},K}(S_{\mathrm{net},K} + H_K D_K^2), \qquad C_{0,K} = e\, D_K H_K L_{\mathrm{net},K} W_{\mathrm{net},K} B_{\mathrm{cov},K},$$

we have, for $0 < \eta \le 1$,

$$\log N(\eta, \mathcal{T}_K, \|\cdot\|_{\infty,\mathrm{TS}}) \le \frac{V_K}{M} \log\left(\frac{C_{0,K}}{\eta}\right). \tag{F.3}$$

Now set

$$\eta_\varepsilon := 1 \wedge \frac{\varepsilon}{4L}.$$

From

$$N(\varepsilon, \mathscr{L}_{K,\lambda}, \|\cdot\|_\infty) \le N\Big(\frac{\varepsilon}{4L}, \mathcal{T}_K, \|\cdot\|_{\infty,\mathrm{TS}}\Big)^M$$

and the monotonicity of covering numbers in the covering radius,

$$\begin{aligned} \log N(\varepsilon, \mathscr{L}_{K,\lambda}, \|\cdot\|_\infty) &\le M \log N(\eta_\varepsilon, \mathcal{T}_K, \|\cdot\|_{\infty,\mathrm{TS}}) \\ &\le V_K \log\left(\frac{C_{0,K}}{\eta_\varepsilon}\right), \end{aligned}$$

where the second inequality follows from (F.3).

Finally,

$$\eta_\varepsilon^{-1} = 1 \vee \frac{4L}{\varepsilon} \le 1 + \frac{4L}{\varepsilon},$$

so

$$\log N(\varepsilon, \mathscr{L}_{K,\lambda}, \|\cdot\|_\infty) \le V_K \log\left[C_{0,K}\left(1 + \frac{4L}{\varepsilon}\right)\right].$$

Evaluating the integral With $B = \mathcal{B}_{K,\lambda}$, let

$$\Gamma_m(K,\lambda) := \inf_{0<\varepsilon_0\le B/2}\left\{\varepsilon_0 + \frac{1}{\sqrt{m}}\int_{\varepsilon_0}^{B/2}\sqrt{\log N(\varepsilon, \mathscr{L}_{K,\lambda}, \|\cdot\|_\infty)}\,d\varepsilon\right\}.$$

Using

$$\log N(\varepsilon, \mathscr{L}_{K,\lambda}, \|\cdot\|_\infty) \le V_K \log\left[C_{0,K}\left(1+\frac{4L}{\varepsilon}\right)\right],$$

and letting $\varepsilon_0 \downarrow 0$, we obtain

$$\Gamma_m(K,\lambda) \le \sqrt{\frac{V_K}{m}}\int_0^{B/2}\sqrt{\log\left[C_{0,K}\left(1+\frac{4L}{\varepsilon}\right)\right]}\,d\varepsilon.$$

Set $b = B/2$ and $\varepsilon = bu$. Since $0 < u \le 1$,

$$1 + \frac{4L}{bu} \le \frac{1+4L/b}{u} = \frac{1+8L/B}{u},$$

and hence

$$\begin{aligned}
\Gamma_m(K,\lambda) &\le b\sqrt{\frac{V_K}{m}}\int_0^1\sqrt{\log\left[C_{0,K}\left(1+\frac{8L}{B}\right)\right] + \log\frac{1}{u}}\,du \\
&\le b\sqrt{\frac{V_K}{m}}\left\{\sqrt{\log\left[C_{0,K}\left(1+\frac{8L}{B}\right)\right]} + \int_0^1\sqrt{\log\frac{1}{u}}\,du\right\} \\
&= \frac{B}{2}\sqrt{\frac{V_K}{m}}\left\{\sqrt{\log\left[C_{0,K}\left(1+\frac{8L}{B}\right)\right]} + \frac{\sqrt{\pi}}{2}\right\},
\end{aligned}$$

where we used $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ and

$$\int_0^1\sqrt{\log(1/u)}\,du = \Gamma(3/2) = \frac{\sqrt{\pi}}{2}.$$

Therefore, substituting this bound into

$$R(\widehat{p}_m) - \inf_{p\in\mathcal{P}^{\mathrm{bd}}_{K,\lambda}} R(p) \le C_{\mathrm{ch}}\Gamma_m(K,\lambda) + B\sqrt{\frac{2}{m}\log\frac{2}{\delta}},$$

gives, with probability at least $1-\delta$,

$$\begin{aligned}
R(\widehat{p}_m) - \inf_{p\in\mathcal{P}^{\mathrm{bd}}_{K,\lambda}} R(p) \le C_{\mathrm{ch}}\frac{\mathcal{B}_{K,\lambda}}{2}\sqrt{\frac{V_K}{m}}&\left\{\sqrt{\log\left[C_{0,K}\left(1+\frac{8L}{\mathcal{B}_{K,\lambda}}\right)\right]} + \frac{\sqrt{\pi}}{2}\right\} \\
&+ \mathcal{B}_{K,\lambda}\sqrt{\frac{2}{m}\log\frac{2}{\delta}}.
\end{aligned}$$

Finally, Theorem 6.2 constructs $p_K \in \mathcal{P}^{\mathrm{bd}}_{K,\lambda}$ with

$$R(p_K) - H(P) \le L\log\frac{1}{1-\lambda} + 2L\rho_K.$$

Therefore

$$R(\widehat{p}_m) - H(P) \le \left[R(\widehat{p}_m) - \inf_{p\in\mathcal{P}^{\mathrm{bd}}_{K,\lambda}} R(p)\right] + R(p_K) - H(P),$$

and combining the preceding estimation and approximation bounds gives Theorem 6.3.

F.1 Proof of Consequences

Proof of Corollary A.5. Write C_{tr} for the constant in Lemma 4.1. Recall that C_{tr} depends only on Q, ν, r, β, τ, while C_{TS} and C_{ch} are universal constants.

Bound the two architecture quantities Since $K \geq 2$ and $Q = M + 2 \geq 1$,

$$H_K = 2K + 1 \leq 2(K+1), \qquad D_K = 2Q(K+1) + 2 \leq 4Q(K+1).$$

Together with $L_{\mathrm{net},K} \leq C_{\mathrm{tr}}(K+1)$, $S_{\mathrm{net},K} \leq C_{\mathrm{tr}}(K+1)^{r+1}$, $W_{\mathrm{net},K} \leq C_{\mathrm{tr}}(K+1)^r$, and

$$B_{\mathrm{cov},K} \leq C_{\mathrm{tr}}(1 \vee R_\lambda)(K+1)^3 \log(K+2),$$

this gives

$$V_K = C_{\mathrm{TS}} M L_{\mathrm{net},K}(S_{\mathrm{net},K} + H_K D_K^2) \leq C_V \overline{V}_K, \tag{F.4}$$

$$C_{0,K} = e\, D_K H_K L_{\mathrm{net},K} W_{\mathrm{net},K} B_{\mathrm{cov},K} \leq C_0 Q(1 \vee R_\lambda)(K+1)^{r+6} \log(K+2), \tag{F.5}$$

where

$$\overline{V}_K = M\left\{(K+1)^{r+2} + Q^2(K+1)^4\right\},$$

and, for example, one may take

$$C_V = C_{\mathrm{TS}} \max\{C_{\mathrm{tr}}^2, 32 C_{\mathrm{tr}}\}, \qquad C_0 = 8e\, C_{\mathrm{tr}}^3.$$

Substitute into Theorem 6.3 Choose a constant

$$C \geq \max\left\{1,\, C_0,\, \frac{C_{\mathrm{ch}}}{2}\sqrt{C_V}\max\left\{1, \frac{\sqrt{\pi}}{2}\right\}\right\},$$

and set

$$\overline{C}_K := CQ(1 \vee R_\lambda)(K+1)^{r+6}\log(K+2).$$

Then (F.5) gives $C_{0,K} \leq \overline{C}_K$, while (F.4) gives $\sqrt{V_K} \leq \sqrt{C_V\, \overline{V}_K}$. Hence the estimation term in (6.6) satisfies

$$C_{\mathrm{ch}}\frac{\mathcal{B}_{K,\lambda}}{2}\sqrt{\frac{V_K}{m}}\left\{\sqrt{\log\left[C_{0,K}\left(1 + \frac{8L}{\mathcal{B}_{K,\lambda}}\right)\right]} + \frac{\sqrt{\pi}}{2}\right\}$$
$$\leq C\mathcal{B}_{K,\lambda}\sqrt{\frac{\overline{V}_K}{m}}\left[1 + \sqrt{\log\left(\overline{C}_K\left(1 + \frac{8L}{\mathcal{B}_{K,\lambda}}\right)\right)}\right].$$

Substituting this bound into (6.6) yields

$$R(\widehat{p}_m) - H(P) \leq L\log\frac{1}{1-\lambda} + 2L\rho_K$$
$$+ C\mathcal{B}_{K,\lambda}\sqrt{\frac{\overline{V}_K}{m}}\left[1 + \sqrt{\log\left(\overline{C}_K\left(1 + \frac{8L}{\mathcal{B}_{K,\lambda}}\right)\right)}\right]$$
$$+ \mathcal{B}_{K,\lambda}\sqrt{\frac{2}{m}\log\frac{2}{\delta}},$$

as claimed.

The constant C depends only on $C_{\mathrm{tr}}, C_{\mathrm{TS}}, C_{\mathrm{ch}}$. Since C_{TS} and C_{ch} are universal and C_{tr} depends only on Q, ν, r, β, τ, the constant C depends only on Q, ν, r, β, τ (equivalently on M, ν, r, β, τ, since $Q = M + 2$). In particular, it is independent of $K, m, L, \lambda, R_\lambda, \delta$, and of the data law P. □

Proof of Proposition A.6. Assume that the constant R_λ in Assumption 5.1 can be chosen uniformly in L. Throughout the proof, $M, r, \beta, \tau, \nu, \lambda, R_\lambda$ are fixed, while L, K, and m may vary. Put $a = \max\{r+2, 4\}$ and $c_\beta = \beta \log 2$.

Approximation term By Eqn (A.8),

$$\frac{\mathfrak{A}_K}{L} = 2C_{\mathrm{tr}} R_\lambda (K+1)^\kappa e^{-c_\beta K}.$$

Hence

$$\frac{\mathfrak{A}_K}{L} = O\big(K^\kappa e^{-c_\beta K}\big), \tag{F.6}$$

where the implicit constant depends only on $Q, \nu, r, \beta, \tau, R_\lambda$, since C_{tr} depends only on Q, ν, r, β, τ. In particular, it is independent of L, K, and m.

Remove the factor L from the finite sample error term Define

$$\overline{\mathcal{B}}_{K,\lambda} := \frac{\mathcal{B}_{K,\lambda}}{L} = \log\frac{M}{\lambda} + \log M + 2\rho_K.$$

Since $\rho_K = C_{\mathrm{tr}} R_\lambda (K+1)^\kappa 2^{-\beta K} \to 0$,

$$\overline{\mathcal{B}}_{K,\lambda} \longrightarrow \overline{\mathcal{B}}_{\infty,\lambda} := \log\frac{M}{\lambda} + \log M > 0.$$

Thus $\overline{\mathcal{B}}_{K,\lambda} = \Theta(1)$, with constants independent of L. Moreover,

$$1 + \frac{8L}{\mathcal{B}_{K,\lambda}} = 1 + \frac{8}{\overline{\mathcal{B}}_{K,\lambda}} = \Theta(1).$$

Using Eqn (A.9),

$$\frac{\mathfrak{E}_{K,m}}{L} = C\overline{\mathcal{B}}_{K,\lambda}\sqrt{\frac{\overline{V}_K}{m}}\left[1 + \sqrt{\log\left(\overline{C}_K\left(1 + \frac{8}{\overline{\mathcal{B}}_{K,\lambda}}\right)\right)}\right]. \tag{F.7}$$

Evaluate the remaining K-dependence By Corollary A.5,

$$\overline{V}_K = M\left\{(K+1)^{r+2} + Q^2(K+1)^4\right\} = \Theta(K^a),$$

and

$$\overline{C}_K = CQ(1 \vee R_\lambda)(K+1)^{r+6}\log(K+2),$$

so $\log\overline{C}_K = \Theta(\log K)$. Together with $\overline{\mathcal{B}}_{K,\lambda} = \Theta(1)$, equation (F.7) therefore gives

$$\frac{\mathfrak{E}_{K,m}}{L} = \Theta\left(\frac{K^{a/2}\sqrt{\log K}}{\sqrt{m}}\right), \qquad K \to \infty. \tag{F.8}$$

The implicit constants depend only on $M, \lambda, R_\lambda, Q, \nu, r, \beta, \tau$ and the universal covering and chaining constants. They are independent of L, K, and m.

Balance approximation and finite sample errors Assume $R_\lambda > 0$. Since both terms contain the same multiplicative factor L, balancing their total values is equivalent to balancing their per-token values. By (F.6) and (F.8), this gives

$$K^\kappa e^{-c_\beta K} \asymp \frac{K^{a/2}\sqrt{\log K}}{\sqrt{m}}.$$

Taking logarithms,

$$c_\beta K = \frac{1}{2}\log m + \left(\kappa - \frac{a}{2}\right)\log K - \frac{1}{2}\log\log K + O(1). \tag{F.9}$$

Equation (F.9) first implies $K = \Theta(\log m)$; hence $\log K = O(\log\log m)$. Substitution back into (F.9) yields

$$K_m = \frac{\log m}{2c_\beta} + O(\log\log m) = \frac{\log m}{2\beta\log 2} + O(\log\log m).$$

The implicit constant is independent of L and m and depends only on the fixed structural quantities listed above.

Finally, $K_m = \Theta(\log m)$ implies

$$\frac{\mathfrak{A}_{K_m}}{L} = \widetilde{O}(m^{-1/2}), \qquad \frac{\mathfrak{E}_{K_m,m}}{L} = \widetilde{O}(m^{-1/2}),$$

uniformly in L. Equivalently, the corresponding total contributions are $L\,\widetilde{O}(m^{-1/2})$. □

G CHANGES UNDER POLYNOMIAL SPATIAL WEIGHTS

Replace the exponential weight in the definition of $\mathcal{B}^{(r)}_{\beta,\tau}$ by a polynomial weight:

$$\|f\|_{\mathcal{B}^{(r)}_{\beta,\tau,\mathrm{poly}}} := \left\| \left((1+\varrho(\gamma))^{\beta} |d_\gamma| \right)_{\mathrm{ord}(\gamma)\le r} \right\|_{\ell^\tau}, \qquad d_\gamma = 0 \text{ if } \mathrm{ord}(\gamma) > r. \tag{G.1}$$

As before, let τ' be the conjugate exponent and set

$$\vartheta := \beta - \frac{r}{\tau'}. \tag{G.2}$$

The argument below requires $\vartheta > 0$, i.e. $\beta > r/\tau'$; for $\tau = 1$ this reduces to $\beta > 0$.

Transformer approximation. The counting bound is unchanged: $n_r(s) \le C_{q,r}(s+1)^{r-1}$. If $\|f\|_{\mathcal{B}^{(r)}_{\beta,\tau,\mathrm{poly}}} \le R$, then for $\tau > 1$,

$$\begin{aligned}
\|f - f_K\|_\infty &\le \sum_{\varrho(\gamma)>K} |d_\gamma| \\
&\le R \left(\sum_{s>K} n_r(s)(s+1)^{-\beta\tau'} \right)^{1/\tau'} \\
&\le CR \left(\sum_{s>K} (s+1)^{r-1-\beta\tau'} \right)^{1/\tau'} \\
&\le CR(K+1)^{-\vartheta}.
\end{aligned} \tag{G.3}$$

For $\tau = 1$,

$$\|f - f_K\|_\infty \le R(K+2)^{-\beta}, \tag{G.4}$$

which is again (G.3) since $\vartheta = \beta$. The same condition $\vartheta > 0$ gives $\sum_\gamma |d_\gamma| \le CR$. Hence one may take $\varepsilon_K = (K+1)^{-\vartheta}$ and $\delta_K \asymp \varepsilon_K$, so that

$$\log \frac{1}{\delta_K} = O(\log(K+2)). \tag{G.5}$$

The product-network construction therefore gives

$$L_{\mathrm{net},K} \lesssim \log(K+2), \qquad W_{\mathrm{net},K} \lesssim (K+1)^r, \qquad S_{\mathrm{net},K} \lesssim (K+1)^r \log(K+2). \tag{G.6}$$

Moreover, $\log \mathrm{Lip}(\widehat{f}_K) \lesssim (\log(K+2))^2$, and choosing the attention scale as in the proof of Theorem 4.1 gives

$$B_{\mathrm{net},K} \lesssim (K+1)^2 (\log(K+2))^2. \tag{G.7}$$

Consequently the analogue of Theorem 4.1 is

$$\sup_{j\in\mathbb{Z}} \|\widehat{F}_{K,j} - F^\circ_j\|_\infty \le C_{\mathrm{tr}} (K+1)^{-\vartheta}, \tag{G.8}$$

with

$$N_K \lesssim (K+1)^r \log(K+2) + q^2 (K+1)^3. \tag{G.9}$$

Here C_{tr} and the implicit constants depend only on q, ν, r, β, τ.

MDLM approximation. Replace Assumption 5.1 by

$$\max_{a\in\mathcal{A}} \|f^\lambda_a\|_{\mathcal{B}^{(r)}_{\beta,\tau,\mathrm{poly}}} \le R_\lambda. \tag{G.10}$$

Applying (G.8) on the padded alphabet gives

$$\max_{\substack{z\in\mathcal{Z}_P,\ i\in\mathcal{M}(z)\\ a\in\mathcal{A}}} \left| \widehat{F}_{a,K,i}(\iota z) - \ell^\lambda_{i,a}(z) \right| \le \rho_K, \qquad \rho_K := C_{\mathrm{tr}} R_\lambda (K+1)^{-\vartheta}, \tag{G.11}$$

where now C_{tr} depends only on Q, ν, r, β, τ. The softmax argument is unchanged:

$$D_{\mathrm{KL}}(\pi_i(\cdot \mid z) \| p_{K,i}(\cdot \mid z)) \le \log \frac{1}{1-\lambda} + 2\rho_K, \tag{G.12}$$

and therefore

$$R(p_K) - H(P) \le L \log \frac{1}{1-\lambda} + 2LC_{\mathrm{tr}} R_\lambda (K+1)^{-\vartheta}. \tag{G.13}$$

Estimation bound. Under (G.6)–(G.7),

$$V_K \lesssim M\left\{(K+1)^r(\log(K+2))^2 + Q^2(K+1)^3\log(K+2)\right\}, \tag{G.14}$$

$$C_{0,K} \lesssim Q(1\vee R_\lambda)(K+1)^{r+4}(\log(K+2))^3. \tag{G.15}$$

Also

$$\frac{\mathcal{B}_{K,\lambda}}{L} = \log\frac{M}{\lambda} + \log M + 2\rho_K = \Theta(1). \tag{G.16}$$

Define

$$a_{\mathrm{poly}} := \max\{r,3\}, \qquad \zeta_r := \begin{cases} 3/2, & r \geq 3, \\ 1, & r < 3. \end{cases} \tag{G.17}$$

Then

$$\frac{\mathfrak{E}_{K,m}}{L} = \Theta\Big(\frac{K^{a_{\mathrm{poly}}/2}(\log K)^{\zeta_r}}{\sqrt{m}}\Big), \qquad \frac{\mathfrak{A}_K}{L} = O(K^{-\vartheta}), \tag{G.18}$$

and Corollary A.5 becomes, up to constants,

$$\frac{R(\widehat{p}_m) - H(P)}{L} \lesssim \log\frac{1}{1-\lambda} + R_\lambda K^{-\vartheta} + \frac{K^{a_{\mathrm{poly}}/2}(\log K)^{\zeta_r}}{\sqrt{m}} + \sqrt{\frac{\log(1/\delta)}{m}}. \tag{G.19}$$

The implicit constants are independent of L, K, m; they depend only on the fixed structural quantities $M, \lambda, R_\lambda, Q, \nu, r, \beta, \tau$ and on the universal covering and chaining constants.

Optimal K Balancing the two K-dependent terms in (G.18) gives

$$K^{-\vartheta} \asymp \frac{K^{a_{\mathrm{poly}}/2}(\log K)^{\zeta_r}}{\sqrt{m}}, \qquad \text{equivalently} \qquad K^{\vartheta + a_{\mathrm{poly}}/2}(\log K)^{\zeta_r} \asymp m^{1/2}. \tag{G.20}$$

Hence

$$K_m = \Theta\Big(m^{1/(2\vartheta+a_{\mathrm{poly}})}(\log m)^{-2\zeta_r/(2\vartheta+a_{\mathrm{poly}})}\Big) = \widetilde{\Theta}\Big(m^{1/(2\vartheta+a_{\mathrm{poly}})}\Big), \tag{G.21}$$

and at this scale

$$\frac{\mathfrak{A}_{K_m}}{L} \asymp \frac{\mathfrak{E}_{K_m,m}}{L} = \widetilde{\Theta}\Big(m^{-\vartheta/(2\vartheta+a_{\mathrm{poly}})}\Big). \tag{G.22}$$

Thus the exponential-weight chain $2^{\beta\varrho(\gamma)} \Rightarrow K^\kappa 2^{-\beta K} \Rightarrow K_m = \Theta(\log m)$ is replaced by

$$(1+\varrho(\gamma))^\beta \quad\Longrightarrow\quad K^{-\vartheta} \quad\Longrightarrow\quad K_m = \widetilde{\Theta}\Big(m^{1/(2\vartheta+a_{\mathrm{poly}})}\Big), \qquad \vartheta = \beta - \frac{r}{\tau'} > 0. \tag{G.23}$$

If $\beta \leq r/\tau'$, then $\sum_{s>K}(s+1)^{r-1-\beta\tau'} = \infty$, so the present argument no longer yields a vanishing uniform truncation error.

H Experimental Setup

We now discuss the experimental setup of Section 7 in detail.

H.1 Ising Model Experiments

We consider a translation-invariant Ising model on sequences

$$X = (X_1, \ldots, X_L) \in \{-1,+1\}^L$$

with distribution

$$P_T(x) = \frac{1}{Z_{T,L}}\exp\left\{\frac{1}{T}\left(\sum_{i=1}^{L}\sum_{r=1}^{R_0} J_r x_i x_{i+r} + h\sum_{i=1}^{L} x_i\right)\right\}, \tag{H.1}$$

where indices are L periodic, so that

$$x_{i+L} = x_i.$$

Here $R_0 < \infty$ is the interaction range, $J_1, \ldots, J_{R_0} \in \mathbb{R}$ are the interaction strengths, $h \in \mathbb{R}$ is the external field, $T > 0$ is the temperature, and $Z_{T,L}$ is the normalizing constant.

H.1.1 Testing the Norm Condition

We test whether the norm condition assumed in Section 5.2.1 holds uniformly in the sequence length. For each odd $L \in \{5, 7, 9, 11, 13\}$, we consider the above model with $J = 1$ and $h = 0$. Writing $c_L := (L-1)/2$, let $i_0 := c_L + 1 = (L+1)/2$ be the central site, which is always masked. We index the remaining coordinates relative to the target by

$$I_{c_L} := \{-c_L, \dots, -1, 1, \dots, c_L\},$$

so each masked context is $z = (z_j)_{j \in I_{c_L}} \in \mathcal{E}^{L-1}$, where $\mathcal{E} = \mathcal{A} \cup \{\mathtt{m}\}$. All 3^{L-1} contexts are enumerated.

For $a \in \mathcal{A}$ and $z = (z_j)_{j \in I_{c_L}} \in \mathcal{E}^{L-1}$, define

$$W_{a,L,T}(z) := \sum_{\substack{x \in \mathcal{A}^L:\, x_{i_0} = a \\ x_{i_0+j} = z_j \text{ whenever } z_j \neq \mathtt{m}}} \exp\left\{ \frac{1}{T} \left(J \sum_{i=1}^{L} x_i x_{i+1} + h \sum_{i=1}^{L} x_i \right) \right\}, \tag{H.2}$$

and compute the posterior exactly as

$$\pi_{i_0,L,T}(a \mid z) = \frac{W_{a,L,T}(z)}{\sum_{b \in \mathcal{A}} W_{b,L,T}(z)}. \tag{H.3}$$

We fix $\lambda = 10^{-3}$ and use the smoothed posterior and logit

$$\pi^{\lambda}_{i_0,L,T}(a \mid z) = (1 - \lambda)\pi_{i_0,L,T}(a \mid z) + \frac{\lambda}{2}, \qquad f^{\lambda}_{a,L,T}(z) = \log \pi^{\lambda}_{i_0,L,T}(a \mid z). \tag{H.4}$$

For every L, we use the same reference measure ν, namely the uniform measure on $\mathcal{E}$, and the same real orthonormal basis of $L^2(\mathcal{E}, \nu)$:

$$\varphi_0 \equiv 1, \qquad \big(\varphi_1(\mathtt{m}), \varphi_1(-1), \varphi_1(+1)\big) = \left(0, -\sqrt{\frac{3}{2}}, \sqrt{\frac{3}{2}}\right), \tag{H.5}$$

$$\big(\varphi_2(\mathtt{m}), \varphi_2(-1), \varphi_2(+1)\big) = \left(-\sqrt{2}, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right). \tag{H.6}$$

Thus, for $\gamma = (\gamma_j)_{j \in I_{c_L}} \in \{0, 1, 2\}^{L-1}$, we define

$$\chi_\gamma(z) = \prod_{j \in \operatorname{supp} \gamma} \varphi_{\gamma_j}(z_j), \qquad \widehat{f}^{\lambda}_{a,L,T}(\gamma) = \frac{1}{3^{L-1}} \sum_{z \in \mathcal{E}^{L-1}} f^{\lambda}_{a,L,T}(z) \chi_\gamma(z). \tag{H.7}$$

As in Section 3, set

$$\mathfrak{w}_\gamma := \|\chi_\gamma\|_\infty, \qquad d^{\lambda}_{a,L,T,\gamma} := \widehat{f}^{\lambda}_{a,L,T}(\gamma) \mathfrak{w}_\gamma, \tag{H.8}$$

where explicitly

$$\mathfrak{w}_\gamma = \prod_{j \in I_{c_L}} \begin{cases} 1, & \gamma_j = 0, \\ \sqrt{3/2}, & \gamma_j = 1, \\ \sqrt{2}, & \gamma_j = 2. \end{cases} \tag{H.9}$$

The spatial radius is

$$\varrho(\gamma) = \begin{cases} \max\{|j| : \gamma_j \neq 0\}, & \gamma \neq 0, \\ 0, & \gamma = 0. \end{cases} \tag{H.10}$$

We fix $\tau = 2$. For each candidate regularity exponent $\beta > 0$, define

$$R_{a,L,T}(\beta) := \left[\sum_{\gamma} 2^{2\beta\varrho(\gamma)} \left| d^{\lambda}_{a,L,T,\gamma} \right|^2 \right]^{1/2}, \qquad R_{L,T}(\beta) := \max_{a \in \mathcal{A}} R_{a,L,T}(\beta). \tag{H.11}$$

We evaluate (H.11) on the parameter grid

$$T \in \left\{ 10, 5, 2, 1, \frac{2}{3} \right\}, \qquad \beta \in \{0, 0.25, 0.50, 0.75, 1.00, 1.25\}, \tag{H.12}$$

for every $L \in \{5, 7, 9, 11, 13\}$. For each pair (T, β), we inspect the curve $L \mapsto R_{L,T}(\beta)$. The regularity assumption is supported when $R_{L,T}(\beta)$ remains bounded as L increases. We also consider

$$G_{L,T}(\beta) := \frac{R_{L+2,T}(\beta)}{R_{L,T}(\beta)}, \tag{H.13}$$

with stabilization corresponding to $G_{L,T}(\beta) \to 1$ as L increases. Thus the experiment tests whether the norm can be bounded by a constant independent of the sequence length over a nontrivial range of temperatures and β.

H.1.2 Testing the condition on $\mathrm{ord}(\gamma)$

We use the same Ising model, posterior logits, reference measure, and Fourier basis as in the preceding experiment. In particular, for $L \in \{5, 7, 9, 11, 13\}$ and $T \in \{10, 5, 2, 1, 2/3\}$, with $J = 1$, $h = 0$, and $\lambda = 10^{-3}$, we compute the exact smoothed posterior logits $f^{\lambda}_{a,L,T}$ for each $a \in \mathcal{A}$ and their normalized Fourier coefficients $d^{\lambda}_{a,L,T,\gamma}$.

For a multi-index γ, recall that $\mathrm{ord}(\gamma) := |\operatorname{supp}(\gamma)|$. To quantify the contribution of interactions of a fixed order, define

$$M_{a,L,T}(k) := \sum_{\mathrm{ord}(\gamma)=k} \left|d^{\lambda}_{a,L,T,\gamma}\right|^2, \qquad k = 1, \dots, L-1, \tag{H.14}$$

and normalize by

$$p_{a,L,T}(k) := \frac{M_{a,L,T}(k)}{\displaystyle\sum_{\mathrm{ord}(\gamma)\geq 1} \left|d^{\lambda}_{a,L,T,\gamma}\right|^2}. \tag{H.15}$$

Thus $\sum_{k=1}^{L-1} p_{a,L,T}(k) = 1$, and $p_{a,L,T}(k)$ measures the fraction of nonconstant coefficient mass carried by interactions involving exactly k coordinates.

More directly, for a candidate interaction-order cutoff r, define the tail

$$E_{a,L,T}(r) := \frac{\displaystyle\sum_{\mathrm{ord}(\gamma)>r} \left|d^{\lambda}_{a,L,T,\gamma}\right|^2}{\displaystyle\sum_{\mathrm{ord}(\gamma)\geq 1} \left|d^{\lambda}_{a,L,T,\gamma}\right|^2} = \sum_{k>r} p_{a,L,T}(k), \tag{H.16}$$

and take the worst case over output symbols,

$$E_{L,T}(r) := \max_{a\in\mathcal{A}} E_{a,L,T}(r). \tag{H.17}$$

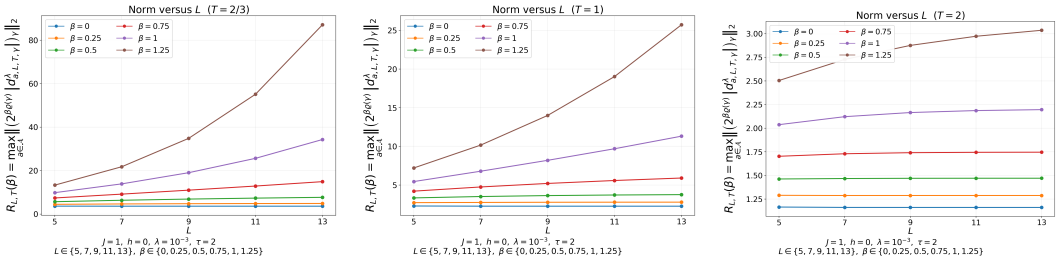
The exact finite-interaction-order assumption would give $E_{L,T}(r) = 0$ beyond some fixed r. We instead test whether $E_{L,T}(r)$ decays rapidly with r.

For each T, we compare the curves $r \mapsto E_{L,T}(r)$ across $L \in \{5, 7, 9, 11, 13\}$ and also report the uniform finite-size envelope

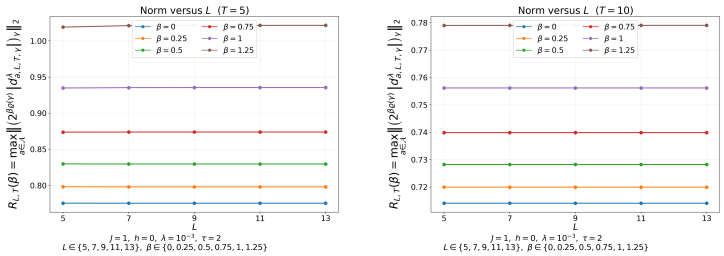
$$E^{\max}_{T}(r) := \max_{L\in\{5,7,9,11,13\}} E_{L,T}(r). \tag{H.18}$$

Rapid decay of $E^{\max}_{T}(r)$ indicates that a fixed, small interaction order captures most of the coefficient mass uniformly over the tested sequence lengths. In particular, approximate linear decay of $\log E^{\max}_{T}(r)$ as a function of r is consistent with an exponential tail $E^{\max}_{T}(r) \approx Ce^{-cr}$.

Figures 4, 5 and 6 summarize all the findings.

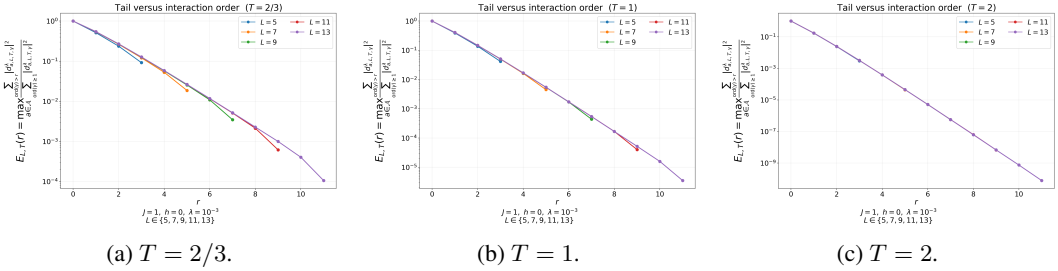


(a) $T = 2/3$. (b) $T = 1$. (c) $T = 2$.

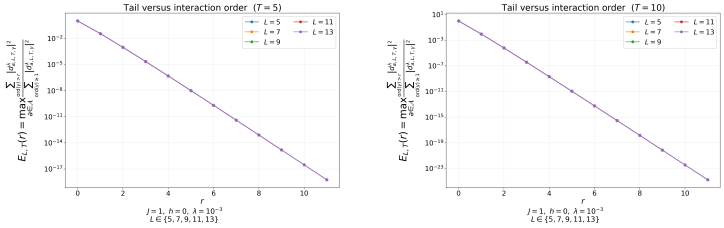


(d) $T = 5$. (e) $T = 10$.

Figure 4: Norm $R_{L,T}(\beta)$ as a function of sequence length L for the five temperatures considered in the Ising-model experiment.

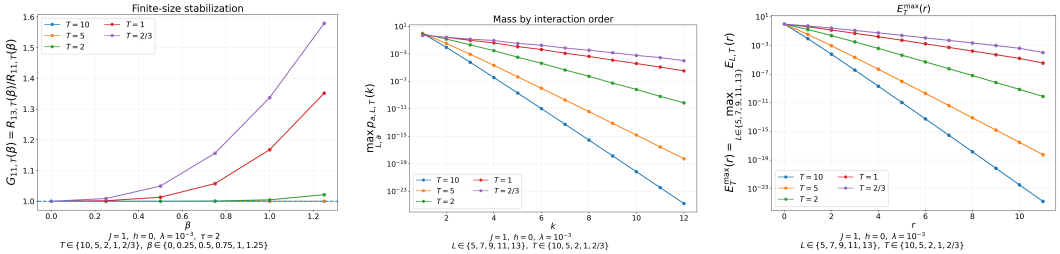


(a) $T = 2/3$. (b) $T = 1$. (c) $T = 2$.



(d) $T = 5$. (e) $T = 10$.

Figure 5: Interaction-order tails $E_{L,T}(r)$ as a function of the cutoff r for the five temperatures considered in the Ising-model experiment. Each panel shows the curves across the tested sequence lengths.



(a) Finite-size ratio $G_{11,T}(\beta)$. (b) Interaction-order mass. (c) Finite-size interaction-order tail.

Figure 6: Aggregate diagnostics for the Ising-model experiments. Left: the finite-size stabilization ratio as a function of β. Center: interaction-order coefficient mass, maximized over the tested sequence lengths. Right: the interaction-order tail, maximized over the tested sequence lengths.

H.2 GTR Model Experiments

The General Time Reversible (GTR) model is the most mathematically flexible, finite-sites, time-reversible continuous-time Markov model used to describe character substitutions. It is used to describe neucleotide sequences (forming DNA and RNA), amino acid sequences (forming proteins), etc. For the experiment, we focus on the 4 state case used for modelling neucleotide sequences. Let

$$\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}.$$

Fix stationary probabilities

$$\pi_{\mathrm{A}}, \pi_{\mathrm{C}}, \pi_{\mathrm{G}}, \pi_{\mathrm{T}} > 0, \qquad \pi_{\mathrm{A}} + \pi_{\mathrm{C}} + \pi_{\mathrm{G}} + \pi_{\mathrm{T}} = 1,$$

and positive exchangeability parameters

$$s_{\mathrm{AC}},\ s_{\mathrm{AG}},\ s_{\mathrm{AT}},\ s_{\mathrm{CG}},\ s_{\mathrm{CT}},\ s_{\mathrm{GT}} > 0.$$

Define

$$g_{\mathrm{A}} := s_{\mathrm{AC}}\pi_{\mathrm{C}} + s_{\mathrm{AG}}\pi_{\mathrm{G}} + s_{\mathrm{AT}}\pi_{\mathrm{T}}, \tag{H.19}$$
$$g_{\mathrm{C}} := s_{\mathrm{AC}}\pi_{\mathrm{A}} + s_{\mathrm{CG}}\pi_{\mathrm{G}} + s_{\mathrm{CT}}\pi_{\mathrm{T}}, \tag{H.20}$$
$$g_{\mathrm{G}} := s_{\mathrm{AG}}\pi_{\mathrm{A}} + s_{\mathrm{CG}}\pi_{\mathrm{C}} + s_{\mathrm{GT}}\pi_{\mathrm{T}}, \tag{H.21}$$
$$g_{\mathrm{T}} := s_{\mathrm{AT}}\pi_{\mathrm{A}} + s_{\mathrm{CT}}\pi_{\mathrm{C}} + s_{\mathrm{GT}}\pi_{\mathrm{G}}, \tag{H.22}$$

and let

$$G := \begin{pmatrix} -g_{\mathrm{A}} & s_{\mathrm{AC}}\pi_{\mathrm{C}} & s_{\mathrm{AG}}\pi_{\mathrm{G}} & s_{\mathrm{AT}}\pi_{\mathrm{T}} \\ s_{\mathrm{AC}}\pi_{\mathrm{A}} & -g_{\mathrm{C}} & s_{\mathrm{CG}}\pi_{\mathrm{G}} & s_{\mathrm{CT}}\pi_{\mathrm{T}} \\ s_{\mathrm{AG}}\pi_{\mathrm{A}} & s_{\mathrm{CG}}\pi_{\mathrm{C}} & -g_{\mathrm{G}} & s_{\mathrm{GT}}\pi_{\mathrm{T}} \\ s_{\mathrm{AT}}\pi_{\mathrm{A}} & s_{\mathrm{CT}}\pi_{\mathrm{C}} & s_{\mathrm{GT}}\pi_{\mathrm{G}} & -g_{\mathrm{T}} \end{pmatrix}. \tag{H.23}$$

Every row of G sums to zero. For a parameter $\eta > 0$, define the one-step transition matrix

$$P_\eta := \exp(\eta G) = \sum_{k=0}^{\infty} \frac{\eta^k G^k}{k!}. \tag{H.24}$$

Thus, for $a, b \in \mathcal{A}$,

$$\mathbb{P}(X_{i+1} = b \mid X_i = a) = [P_\eta]_{ab}. \tag{H.25}$$

The chain is initialized according to

$$\mathbb{P}(X_1 = a) = \pi_a, \qquad a \in \mathcal{A}. \tag{H.26}$$

Consequently, for any sequence $x = (x_1, \ldots, x_L) \in \mathcal{A}^L$,

$$P_{\eta,L}(x) = \pi_{x_1} \prod_{i=1}^{L-1} [P_\eta]_{x_i x_{i+1}}. \tag{H.27}$$

The symmetry $s_{ab} = s_{ba}$ implies $\pi_a G_{ab} = \pi_b G_{ba}$ for every $a, b \in \mathcal{A}$, and hence $\pi_a [P_\eta]_{ab} = \pi_b [P_\eta]_{ba}$. In particular, $(\pi_{\mathrm{A}}, \pi_{\mathrm{C}}, \pi_{\mathrm{G}}, \pi_{\mathrm{T}})$ is a stationary distribution of the chain.

H.2.1 Testing the Norm Condition

We consider the alphabet $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ with stationary distribution $\pi = (0.3, 0.2, 0.2, 0.3)$. We set $s_{\mathrm{AG}} = s_{\mathrm{CT}} = 4$ and $s_{\mathrm{AC}} = s_{\mathrm{AT}} = s_{\mathrm{CG}} = s_{\mathrm{GT}} = 1$, giving

$$G = \begin{pmatrix} -1.3 & 0.2 & 0.8 & 0.3 \\ 0.3 & -1.7 & 0.2 & 1.2 \\ 1.2 & 0.2 & -1.7 & 0.3 \\ 0.3 & 0.8 & 0.2 & -1.3 \end{pmatrix}. \tag{H.28}$$

For $\eta \in \{0.05, 0.1, 0.25, 0.5, 1, 2\}$, let $P_\eta = \exp(\eta G)$ and define, for $x = (x_1, \ldots, x_L) \in \mathcal{A}^L$,

$$P_{\eta,L}(x) = \pi_{x_1} \prod_{i=1}^{L-1} [P_\eta]_{x_i x_{i+1}}. \tag{H.29}$$

We use sequence lengths $L \in \{3, 5, 7, 9, 11\}$. Writing $c_L := (L-1)/2$, the target position is $i_0 = c_L + 1$, and the remaining positions are indexed relative to the target by $I_{c_L} := \{-c_L, \ldots, -1, 1, \ldots, c_L\}$.

Let $\mathcal{E} = \mathcal{A} \cup \{\mathtt{m}\}$, and let $z = (z_j)_{j \in I_{c_L}} \in \mathcal{E}^{2c_L}$ denote a partially observed context, with the target position fixed to be masked. Define $O(z) := \{j \in I_{c_L} : z_j \neq \mathtt{m}\}$. For $a \in \mathcal{A}$, the exact clean-data posterior is

$$\pi_{L,\eta}(a \mid z) := \frac{\displaystyle\sum_{\substack{x \in \mathcal{A}^L: \\ x_{i_0} = a,\ x_{i_0+j} = z_j\ \forall j \in O(z)}} P_{\eta,L}(x)}{\displaystyle\sum_{\substack{x \in \mathcal{A}^L: \\ x_{i_0+j} = z_j\ \forall j \in O(z)}} P_{\eta,L}(x)}. \tag{H.30}$$

With $\lambda = 10^{-3}$, define the smoothed posterior and logit by

$$\pi^\lambda_{L,\eta}(a \mid z) := (1-\lambda)\pi_{L,\eta}(a \mid z) + \frac{\lambda}{4}, \qquad f^\lambda_{a,L,\eta}(z) := \log \pi^\lambda_{L,\eta}(a \mid z). \tag{H.31}$$

We take the reference measure ν to be uniform on $\mathcal{E}$, so that $\nu(e) = 1/5$ for all $e \in \mathcal{E}$. In the ordering $(\mathtt{m}, \mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T})$, we use the orthonormal basis

$$\varphi_0 = (1,1,1,1,1), \qquad \varphi_1 = \sqrt{\frac{5}{2}}(1,-1,0,0,0), \tag{H.32}$$

$$\varphi_2 = \sqrt{\frac{5}{6}}(1,1,-2,0,0), \qquad \varphi_3 = \sqrt{\frac{5}{12}}(1,1,1,-3,0), \tag{H.33}$$

$$\varphi_4 = \frac{1}{2}(1,1,1,1,-4), \tag{H.34}$$

which satisfies $\frac{1}{5}\sum_{e \in \mathcal{E}} \varphi_u(e)\varphi_v(e) = \mathbf{1}\{u = v\}$.

For $\gamma = (\gamma_j)_{j \in I_{c_L}} \in \{0,1,2,3,4\}^{2c_L}$, define $\chi_\gamma(z) := \prod_{j \in I_{c_L}} \varphi_{\gamma_j}(z_j)$ and $\mathfrak{w}_\gamma := \|\chi_\gamma\|_\infty$. The Fourier coefficients and their normalized versions are

$$\widehat{f}^\lambda_{a,L,\eta}(\gamma) := \mathbb{E}_{Z \sim \nu^{\otimes 2c_L}}\left[f^\lambda_{a,L,\eta}(Z)\chi_\gamma(Z)\right] = \frac{1}{5^{2c_L}} \sum_{z \in \mathcal{E}^{2c_L}} f^\lambda_{a,L,\eta}(z)\chi_\gamma(z), \qquad d^\lambda_{a,L,\eta,\gamma} := \widehat{f}^\lambda_{a,L,\eta}(\gamma)\mathfrak{w}_\gamma. \tag{H.35}$$

We define

$$\varrho(\gamma) := \begin{cases} \max\{|j| : j \in I_{c_L},\ \gamma_j \neq 0\}, & \gamma \neq 0, \\ 0, & \gamma = 0. \end{cases} \tag{H.36}$$

Fixing $\tau = 2$, for each $\beta \in \{0, 0.25, 0.5, 0.75, 1, 1.25\}$ we compute

$$R_{L,\eta}(\beta) := \max_{a \in \mathcal{A}} \left[\sum_{\gamma \in \{0,1,2,3,4\}^{2c_L}} 2^{2\beta\varrho(\gamma)} \left|d^\lambda_{a,L,\eta,\gamma}\right|^2\right]^{1/2}. \tag{H.37}$$

For every pair (η, β), we plot $L \mapsto R_{L,\eta}(\beta)$ over $L \in \{3, 5, 7, 9, 11\}$. As a finite-size stabilization diagnostic, we additionally compute

$$G_{9,\eta}(\beta) := \frac{R_{11,\eta}(\beta)}{R_{9,\eta}(\beta)}. \tag{H.38}$$

Values of $G_{9,\eta}(\beta)$ close to one indicate stabilization of the norm over the largest sequence lengths considered.

H.2.2 Testing the condition on $\mathrm{ord}(\gamma)$

The interaction order of γ is $\mathrm{ord}(\gamma) := |\{j \in I_{c_L} : \gamma_j \neq 0\}|$. For each $k = 1, \ldots, L-1$, define the normalized coefficient mass at order k by

$$p_{a,L,\eta}(k) := \frac{\displaystyle\sum_{\mathrm{ord}(\gamma)=k} |d^{\lambda}_{a,L,\eta,\gamma}|^2}{\displaystyle\sum_{\mathrm{ord}(\gamma)\geq 1} |d^{\lambda}_{a,L,\eta,\gamma}|^2}. \tag{H.39}$$

At $L = 11$, we plot $k \mapsto \max_{a\in\mathcal{A}} p_{a,11,\eta}(k)$ for all six values of η.

To measure the contribution of interactions above a candidate cutoff r, define

$$E_{a,L,\eta}(r) := \frac{\displaystyle\sum_{\mathrm{ord}(\gamma)>r} |d^{\lambda}_{a,L,\eta,\gamma}|^2}{\displaystyle\sum_{\mathrm{ord}(\gamma)\geq 1} |d^{\lambda}_{a,L,\eta,\gamma}|^2}, \qquad E_{L,\eta}(r) := \max_{a\in\mathcal{A}} E_{a,L,\eta}(r), \tag{H.40}$$

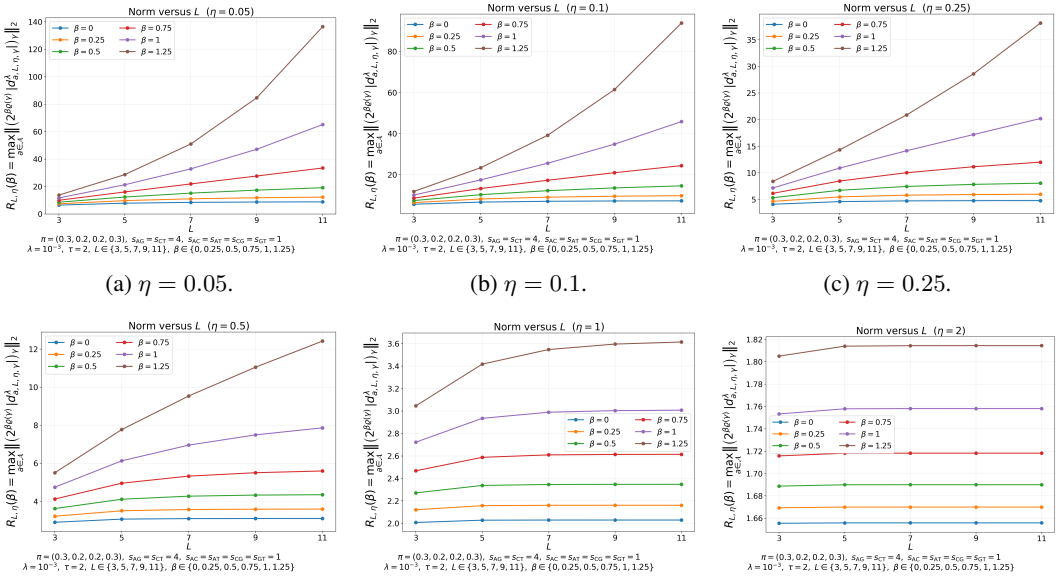
for $r = 0, \ldots, L-1$. For each $\eta \in \{0.05, 0.1, 0.25, 0.5, 1, 2\}$, we plot $r \mapsto E_{L,\eta}(r)$ for $L \in \{3, 5, 7, 9, 11\}$.

Finally, define the finite-size envelope

$$E^{\max}_{\eta}(r) := \max_{L\in\{3,5,7,9,11\}} E_{L,\eta}(r), \tag{H.41}$$

where $E_{L,\eta}(r) := 0$ for $r \geq L-1$. We plot $r \mapsto E^{\max}_{\eta}(r)$ for $r = 0, \ldots, 10$ and all six values of η. To examine the decay rate, we additionally plot $r \mapsto \log_{10} E^{\max}_{\eta}(r)$ for $r = 0, \ldots, 9$, omitting the terminal value $E^{\max}_{\eta}(10) = 0$.

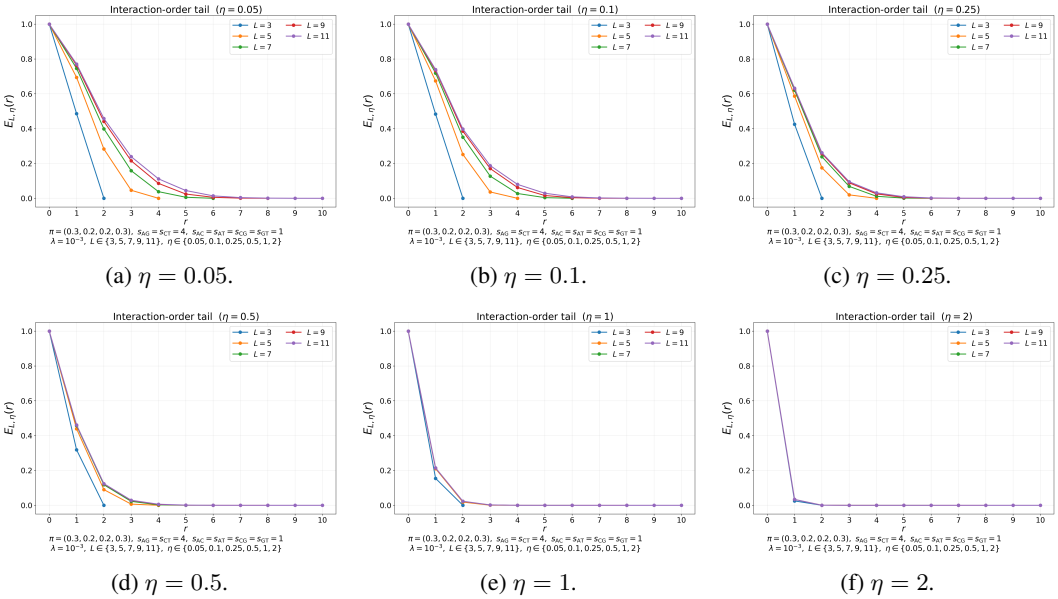
Figures 7, 8, 9 summarize all the findings.



(a) $\eta = 0.05$. (b) $\eta = 0.1$. (c) $\eta = 0.25$.

(d) $\eta = 0.5$. (e) $\eta = 1$. (f) $\eta = 2$.

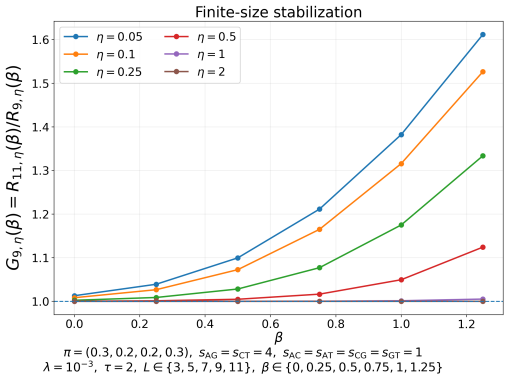
Figure 7: Norm $R_{L,\eta}(\beta)$ as a function of sequence length L for the six values of η considered in the GTR-model experiment.



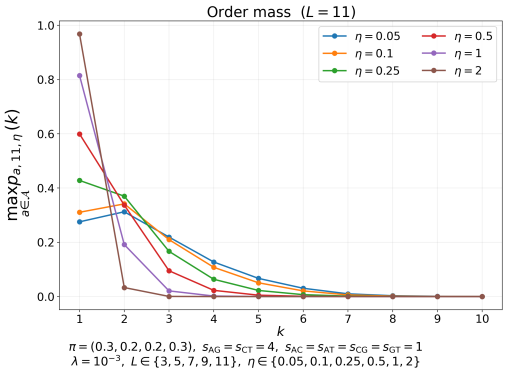
(a) $\eta = 0.05$. (b) $\eta = 0.1$. (c) $\eta = 0.25$.

(d) $\eta = 0.5$. (e) $\eta = 1$. (f) $\eta = 2$.

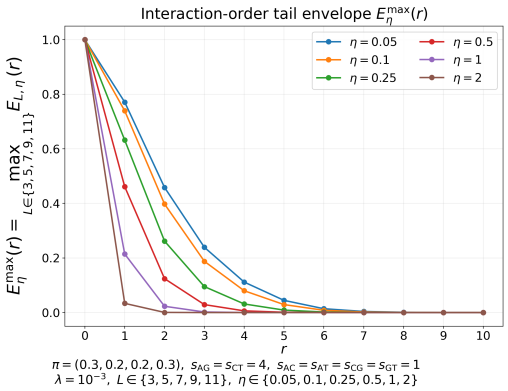
Figure 8: Interaction-order tails $E_{L,\eta}(r)$ as a function of the cutoff r for the six values of η considered in the GTR-model experiment. Each panel shows the curves across the tested sequence lengths.



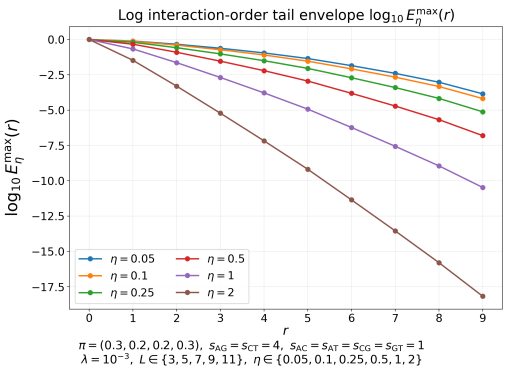
(a) Finite-size ratio $G_{9,\eta}(\beta)$.



(b) Interaction-order mass at $L = 11$.



(c) Finite-size interaction-order tail envelope.



(d) Logarithmic interaction-order tail envelope.

Figure 9: Aggregate diagnostics for the GTR-model experiments. Top left: the finite-size stabilization ratio as a function of β. Top right: normalized coefficient mass by interaction order at $L = 11$. Bottom left: the finite-size interaction-order tail envelope. Bottom right: the same envelope on a base-10 logarithmic scale.