TRANSFORMER-BASED MASKED DIFFUSION MODELS PROVABLY LEARN DISTRIBUTIONS WITH LATENT STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked diffusion models (MDMs) generate discrete data by iteratively unmasking multiple tokens in parallel using a learned mask predictor. Their output carries two errors: parallelization error due to parallel sampling, and estimation error from the mask predictor learned with finite samples. Existing theory treats these errors separately, and the available statistical bounds typically scale with the number V^L of length-L sequences over a vocabulary of size V, suffering from the curse of dimensionality. In this work, we establish end-to-end statistical guarantees showing that transformer-based MDMs learn structured distributions with a sample complexity governed by their latent structure. Specifically, we model this structure through mixtures of hidden Markov trees, which capture latent dependencies within components and heterogeneity across components. When the data distribution is well approximated by such a mixture, a transformer-based MDM with parallel unmasking schedules achieves ε-accurate sampling in KL divergence with $\widetilde{O}(LK_{\mathrm{par}}/\varepsilon)$ training samples, where K_{par} is polynomial in L, V, the number of hidden states, and the number of mixture components. The transformer network and sampling procedure require no knowledge of the tree topologies or component parameters, thereby adapting to the unknown latent structure automatically. Overall, our results provide theoretical justification for the empirical effectiveness of MDMs on real-world data.

1 INTRODUCTION

Masked diffusion models (MDMs) (Austin et al., 2021; Campbell et al., 2022) have emerged as a powerful generative modeling paradigm for discrete data, with applications including language models (Sahoo et al., 2024; Shi et al., 2024), code generation (Gong et al., 2026) and protein sequence learning (Wang et al., 2024). Inspired by diffusion models on continuous spaces (Ho et al., 2020; Song et al., 2020), an MDM generates samples from a data distribution $p^\star$ by learning to reverse a categorical corruption process over a discrete state space. Concretely, an MDM augments the vocabulary with an absorbing *mask* token[1]. The forward process gradually masks a sequence drawn from $p^\star$ until it is fully masked, and the reverse process starts from a fully masked sequence and iteratively unmasks a subset of positions.

Ideally, each unmasking step would sample the selected tokens jointly, conditioned on the revealed context. Since learning these joint conditionals is computationally prohibitive, practical MDMs learn conditional marginals instead and sample from their product (Nie et al., 2026). A transformer-based *mask predictor*, trained on finite samples from $p^\star$, can compute these conditional marginals for all masked positions in a single forward pass. This enables parallel token generation and breaks the sequential bottleneck of autoregressive models. However, sampling tokens independently given the same context ignores dependencies among simultaneously generated tokens. The output of an MDM thus carries two errors: a *parallelization error*, which persists even with exact marginals, and an *estimation error*, because the mask predictor is learned from finite samples.

[1] We use "tokens" to refer to discrete states for convenience; our framework and results apply to any discrete data beyond text.

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.

Despite the empirical success of MDMs, their end-to-end sampling guarantees remain largely unexplored, as existing theory typically treats training and sampling separately. One line of work takes the mask predictor as given, either exact or with a prescribed error. Li & Cai (2025) derived the sampling error bound for parallel generation with a fixed-cardinality unmasking schedule via information-theoretic measures, and Zhao & Cai (2026); Dmitriev et al. (2026); Wainwright (2026) developed fixed-cardinality unmasking schedules that exploit intrinsic data complexity. In addition, Cai & Li (2026) established the sampling guarantees of confidence-based unmasking. These works do not characterize the statistical cost of learning the mask predictor. Another line of work (Wakasugi & Suzuki, 2026; Zhang et al.) studies finite-sample estimation for the mask predictor, but either restricts attention to one-token-at-a-time unmasking or does not consider the transformer architectures that enable parallel sampling in practice. Moreover, these bounds scale with the size V^L of the sequence space or of the support of $p^\star$, where V is the vocabulary size and L is the sequence length. Thus, they suffer from the curse of dimensionality and offer limited insight into the empirical success of MDMs on structured data.

Filling this gap requires an end-to-end analysis that connect the statistical cost of learning a transformer-based mask predictor to the accuracy of parallel generation, with guarantees that capture latent structure in the data. The dependencies observed in text and biological sequences (Chiu & Rush, 2020; Frazer et al., 2021) motivate studying distributions with compact latent representations. Such distributions can nevertheless assign positive probability to all V^L sequences, so bounds governed by the support size can remain exponential in L, obscuring the potential statistical benefits of latent structure. This leads to our central question:

How many training samples does a transformer-based MDM need to sample accurately from $p^\star$, and how does this number depend on the latent structure of the data?

1.1 Main contribution

In this work, we establish end-to-end statistical guarantees for transformer-based MDMs, connecting the finite-sample learning of the mask predictor to the KL accuracy of the generated distribution.

To overcome the curse of dimensionality, we model low-dimensional structure by mixtures of hidden Markov trees, which capture structured latent dependence within each component and heterogeneity across components. In particular, the data distribution need not belong to this class exactly.

When the data distribution is approximated by such a model, we show that a transformer-based MDM with a parallel unmasking schedule achieves ε-accurate sampling in KL divergence with

$$\widetilde{O}(LK_{\mathrm{par}}/\varepsilon) \tag{1}$$

training samples. Here, L is the sequence length and the structural complexity K_{par} is polynomial in the sequence length, vocabulary size V, number of hidden states, and number of components, rather than scaling with the size V^L of the sequence space. Moreover, the transformer architecture for the mask predictor and the parallel sampling procedure require no prior knowledge of the tree topologies or component parameters, so the MDM adapts to the unknown latent structure of the data.

To the best of our knowledge, these are the first end-to-end statistical guarantees for transformer-based MDMs that cover both mask-predictor estimation and parallel unmasking while adapting to latent structure, providing theoretical support for the effectiveness of MDMs on structured data.

1.2 Related work

Sample complexity for diffusion models. Continuous diffusion models admit statistical guarantees under density smoothness (Oko et al., 2023; Cai & Li, 2025) and low-dimensional structure (Azangulov et al., 2024; Wu & Cai, 2026). For discrete diffusion, Wakasugi & Suzuki (2026) establish state-size-independent guarantees under smoothness and spectral assumptions after continuous embedding, but their symmetric-generator assumption excludes MDMs. Srikanth et al. (2025) establish finite-sample KL guarantees under bounded-score and optimization regularity assumptions. Zhang et al. derive support-dependent bounds for masking using empirical kernel-based estimators, and Cho & Wu (2026) establish near-minimax estimation and sampling guarantees with full state-space dependence.

Convergence theory for MDMs. Even with exact conditional marginals, parallel unmasking introduces error from independent token sampling. Li & Cai (2025) and Chen et al. (2025) characterize this error using information-theoretic measures, while Zhao & Cai (2026) and Dmitriev et al. (2026) develop schedules that adapt to intrinsic data complexity. Cai & Li (2026) analyze entropy-based confidence decoding with exact conditional marginals. Complementary work studies continuous-time formulations and discretization error Chen & Ying (2024); Liang et al. (2026). These analysis assume exact predictors or prescribed prediction errors without quantifying their finite-sample learning cost.

1.3 Notation

For $a, b \in \mathbb{R}$, write $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. For a positive integer M, let $[M] := \{1, \cdots, M\}$. Let $\mathbf{1}\{\cdot\}$ denote the indicator function. For $x \in \mathbb{R}^d$, define $\|\cdot\|_p$ as its p-norm, i.e., $\|x\|_p := (\sum_{i=1}^d |x_i|^p)^{1/p}$, and write $\|x\|_\infty := \max_i |x_i|$. We further define $\|x\|_0$ by $\|x\|_0 := \sum_{i\in[d]} \mathbf{1}\{x_i \neq 0\}$. For $W \subseteq [d]$, let $x_W := (x_i)_{i\in W}$. For any matrix $A \in \mathbb{R}^{m\times n}$, denote its i-th row, j-th column and Frobenius norm by $A_{i,\cdot}$, $A_{\cdot,j}$ and $\|A\|_F$, respectively. We denote by $\mathsf{supp}(p)$ the support of probability distribution p. We use $\mathrm{poly}\log(\cdot)$ to denote an unspecified polynomial in the logarithm of its arguments. For positive functions f, g, write $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ if $f(n) \leq Cg(n)$ for some universal constant $C > 0$. The notation $\widetilde{O}$ denotes the respective bounds up to some logarithmic factors.

2 Preliminaries

2.1 Masked diffusion models

Let $\mathbb{V}$ denote the vocabulary set with $|\mathbb{V}| = V$, and let $L \in \mathbb{N}_+$ be the sequence length. Denote $\mathtt{M} \notin \mathbb{V}$ as the mask symbol, and define the extended vocabulary $\overline{\mathbb{V}} := \mathbb{V} \cup \{\mathtt{M}\}$. For convenience, we refer to elements of $\mathbb{V}$ as tokens, although our framework applies to discrete sequences beyond text.

Forward masking process. Given token sequence $X = (X_1, \ldots, X_L) \sim p^\star$ on $\mathbb{V}^L$, consider a fixed-cardinality masking schedule $s = (s_1, \ldots, s_T)$ satisfying $s_t \in \mathbb{N}_+$ and $\sum_{t=1}^T s_t = L$. Initialize $M^0 = \emptyset$. At each step $t \in [T]$, form a random set S^t by selecting s_t positions uniformly from $[L] \setminus M^{t-1}$ and then update by $M^t = M^{t-1} \cup S^t$. The sequence X at time step t is corrupted by replacing the tokens at positions in M^t with $\mathtt{M}$, and leaves other tokens unchanged. After step T, we have a fully masked sequence $X = (\mathtt{M}, \ldots, \mathtt{M})$ with $M^T = [L]$.

Train the mask predictor. For masked diffusion models, concrete score is decoupled into a known time-dependent term and a time-independent clean-data conditional distribution Ou et al. (2025). For a masked sequence $x \in \overline{\mathbb{V}}^L$, let $W(x) := \{i \in [L] : x_i \neq \mathtt{M}\}$ be the index set of unmasked positions. The target conditional distributions are

$$p_i^\star(v \mid x) := \mathbb{P}_{X\sim p^\star}\big[X_i = v \mid X_{W(x)} = x_{W(x)}\big], \quad \forall v \in \mathbb{V}, \forall i \in [L] \setminus W(x). \tag{2}$$

For training, draw $\tau \in [T]$ with $\mathbb{P}[\tau = t] := s_t/L$, sample M^τ uniformly among subsets of $[L]$ of size $\sum_{t=1}^\tau s_t$ and set $W^\tau := [L] \setminus M^\tau$. A mask predictor $p_{\theta,i}$ is trained via (Shi et al., 2024; Sahoo et al., 2024)

$$\min_\theta \overline{R}(p_\theta) := \mathbb{E}_{X\sim p^\star}[\ell_{p_\theta}(X)] \quad \text{with} \quad \ell_q(x) := \mathbb{E}_{\tau, M^\tau}\Big[\frac{1}{|M^\tau|}\sum_{i\in M^\tau} -\log q_i(x_i \mid x_{W^\tau})\Big], \tag{3}$$

defined for any conditional mask predictor $q = (q_i)_{i=1}^L$. Here x_{W^τ} is the corrupted sequence of x that replaces tokens outside W^τ with $\mathtt{M}$. Define the excess risk $\mathcal{L}(\theta) := \overline{R}(p_\theta) - \overline{R}(p^\star)$, where $\overline{R}(p^\star)$ denotes the population loss evaluated at the true conditional marginals in (2). Minimizing $\mathcal{L}(\theta)$ is equivalent to minimizing $\overline{R}(p_\theta)$. As shown in Lemma 4 in Appendix B.1, $\mathcal{L}(\theta)$ corresponds to the KL divergence between the true and predicted conditional distributions. Given n i.i.d. training samples $\{X^{(j)}\}_{j=1}^n \sim p^\star$, define the empirical risk $\widehat{R}_n(p_\theta) := \frac{1}{n}\sum_{j=1}^n \ell_{p_\theta}(X^{(j)})$ and empirical risk minimizer (ERM) $\widehat{\theta} \in \arg\min_\theta \widehat{R}_n(p_\theta)$. In practice, the masking expectation in (3) is estimated by sampling times and masks.

Sampling procedure. Given a mask predictor p_θ and a fixed-cardinality unmasking schedule $s = (s_1, \dots, s_T)$, initialize the sequence $Y = (\mathtt{M}, \dots, \mathtt{M})$ and the set of unmasked positions $W^T = \emptyset$. For iteration $t = T, \dots, 1$, draw an random unmasking set S^t uniformly among the subsets of $[L] \setminus W^t$ with size s_t, set $W^{t-1} = W^t \cup S^t$, sample *simultaneously* and *independently*

$$Y_i \sim p_{\theta,i}(\cdot \mid Y_{W^t}) \quad \text{for } i \in S^t, \tag{4}$$

and keep other tokens unchanged. Since $\sum_{t=1}^{T} s_t = L$, the final iterate $Y \in \mathbb{V}^L$ is fully unmasked and is the generated sample. We write $p_{Y|S}$ for its distribution to emphasize its dependence on the random unmasking sets $S = (S^1, \dots, S^T)$.

2.2 Transformer structure for estimation

We specify the transformer structure Vaswani et al. (2017) used for learning the mask predictor.

Transformer block. Each layer maps $Z \in \mathbb{R}^{L\times D}$ to $\mathbb{R}^{L\times D}$, with each row containing the D-dimensional representation. A single head is defined by $\mathrm{head}\big(Z; Q, K, V\big) := \frac{1}{L}\sigma_{\mathrm{attn}}\big((ZQ)(ZK)^\top\big)ZV$ with sigmoid activation $\sigma_{\mathrm{attn}}(x) := \frac{1}{1+e^{-x}}$ applied entrywise. Here $Q, K, V \in \mathbb{R}^{D\times D}$ denote query, key and value matrices, respectively. Combining M heads with a residual connection gives:

$$\mathsf{Attn}_{\theta_{\mathrm{attn}}}(Z) := Z + \sum_{m=1}^{M} \mathrm{head}\big(Z; Q_m, K_m, V_m\big), \quad \text{with } \theta_{\mathrm{attn}} := \big(Q_m, K_m, V_m\big)_{m\in[M]} \tag{5}$$

The rowwise feed-forward layer satisfies $\mathsf{FFN}_{\theta_{\mathrm{ff}}}(Z) := Z + \mathrm{ReLU}\left(ZW\right)U$ with $W \in \mathbb{R}^{D\times D_{\mathrm{ff}}}, U \in \mathbb{R}^{D_{\mathrm{ff}}\times D}$, and $\mathrm{ReLU}(x) := x \vee 0$ applied entrywise. Parameters $\theta_{\mathrm{ff}} := (W, U)$. A transformer block is then defined by:

$$\mathsf{T}_\theta(Z) := \mathsf{FFN}_{\theta_{\mathrm{ff}}}\big(\mathsf{Attn}_{\theta_{\mathrm{attn}}}(Z)\big), \quad \text{with } \theta := (\theta_{\mathrm{attn}}, \theta_{\mathrm{ff}}), \tag{6}$$

and preserves the $L \times D$ dimensions.

Sparse transformer architecture for mask predictor. We build the mask predictor by composing an input embedding, N transformer blocks, an output embedding and a clipped rowwise softmax.

Input embedding. For a masked sequence $x \in \overline{\mathbb{V}}^L$, let $E \in \mathbb{R}^{(V+1)\times D}$ be its input embedding matrix, whose rows give the representations of each token $v \in \overline{\mathbb{V}}$ in $\mathbb{R}^D$ with $V = |\mathbb{V}|$. Let $P \in \mathbb{R}^{L\times D}$ be the positional embedding matrix. The input embedding is defined by

$$\big(\mathsf{Embed}_{E,P}(x)\big)_{i,\cdot} := E_{x_i,\cdot} + P_{i,\cdot}, \quad \forall i \in [L]. \tag{7}$$

Both matrices are trainable. We control the magnitude of the embedding vectors by some chosen parameter $R_{\mathrm{embed}} > 0$, i.e., $\max_{i\in[L], v\in\overline{\mathbb{V}}} \|E_{v,\cdot} + P_{i,\cdot}\|_2 \le R_{\mathrm{embed}}$.

N transformer blocks. For l-th transformer block, collect its parameters as $\theta_{\mathrm{tf}}^{(l)} := \big((Q_{l,m}, K_{l,m}, V_{l,m})_{m=1}^M, W_l, U_l\big)$, and denote the corresponding block by $\mathsf{T}^{(l)} := \mathsf{T}_{\theta_{\mathrm{tf}}^{(l)}}$. We successively apply N transformer blocks in (6):

$$\mathsf{TF}_{\theta_{\mathrm{tf}}}(Z) := \mathsf{T}^{(N)} \circ \cdots \circ \mathsf{T}^{(1)}(Z), \text{ where } \theta_{\mathrm{tf}} := (\theta_{\mathrm{tf}}^{(1)}, \cdots \theta_{\mathrm{tf}}^{(N)}).$$

Each block perserves the output dimension.

Output embedding. We map the final hidden representation in $\mathbb{R}^{L\times D}$ to vocabulary logits $\mathbb{R}^{L\times V}$ using the output embedding matrix $W_{\mathrm{out}} \in \mathbb{R}^{D\times V}$:

$$\mathsf{OutEmbed}_{W_{\mathrm{out}}}(Z) := ZW_{\mathrm{out}} \in \mathbb{R}^{L\times V}.$$

Clipped rowwise softmax. For a fixed chosen $G > 0$, we clip each logit entrywise to $[-G, 0]$, i.e., $\mathrm{Clip}_{[-G,0]}(x) := \max\{-G, \min\{x, 0\}\}$, and apply softmax separately to each row:

$$\big(\mathsf{Softmax}_G(Z)\big)_{i,\cdot} := \mathrm{softmax}\big(\mathrm{Clip}_{[-G,0]}(Z_{i,\cdot})\big), \quad \text{with } \mathrm{softmax}(x) := \Big(\frac{e^{x_i}}{\sum_j e^{x_j}}\Big)_i. \tag{8}$$

Parameter and function class. Combining these operations gives our mask predictor:

$$\widetilde{\mathsf{TF}}_\theta(x) = \mathsf{Softmax}_G\big(\mathsf{OutEmbed}_{W_{\mathrm{out}}}\big(\mathsf{TF}_{\theta_{\mathrm{tf}}}\big(\mathsf{Embed}_{E,P}(x)\big)\big)\big). \tag{9}$$

Let $\Theta := \Theta(D, N, M, D_{\mathrm{ff}}, B, S; R_{\mathrm{embed}}, G)$ consists of parameters $\theta := \big(E, P, \theta_{\mathrm{tf}}, W_{\mathrm{out}}\big)$ satisfying the embedding bound above, with each trainable matrix having Frobenius norm at most B and at most S nonzero parameters in total.

The predictor returns a probability distribution at each masked position i in a single forward pass:

$$p_{\theta,i}(\cdot \mid x) := \big(\widetilde{\mathsf{TF}}_\theta(x)\big)_{i,\cdot} \tag{10}$$

We define its approximation error by $\varepsilon_{\mathrm{approx}}(\Theta) := \min_{\theta\in\Theta} \mathcal{L}(\theta)$.

3 Problem setup: mixtures of hidden Markov trees

Without structural assumptions, a distribution on $\mathbb{V}^L$ has $V^L - 1$ parameters, and learning it can require a number of samples that grows with V^L. To exploit low-dimensional structure, we introduce mixtures of hidden Markov trees (MHMTs) as an approximation class for $p^\star$, enabling statistical guarantees governed by the complexity of its latent representation.

A *hidden Markov tree* (HMT) is a latent-variable model that captures dependency structure (Crouse et al., 1998), extending a hidden Markov model (HMM) (Baum & Petrie, 1966) from a chain to a tree. An HMM associates each token position $i \in [L]$ with a hidden state $Z_i \in [J]$, where J is the cardinality of hidden space. The states form a Markov chain $Z_1 \to \cdots \to Z_L$, and, conditioned on these states, the tokens are generated independently, with each X_i depending only on Z_i. An HMT allows the latent states to evolve along a tree instead of a chain, supporting more general dependence structures. And a *mixture of hidden Markov trees* (MHMT) further allows the tree topology and parameters to vary across components.

MHMTs contains mixture of HMMs, which captures sequential dependence and diversity in language models (Ypma & Heskes, 2002; Chiu & Rush, 2020; Zekri et al., 2024). In addition, they are widely used to capture biological dependencies in protein-DNA binding sites (Barash et al., 2003), gene modules (Li et al., 2019), and heterogeneous cell lineages (Mohammadi et al., 2022).

We now introduce the formal definition of MHMTs, following (Meila & Jordan, 2000).

Definition 1 (Mixture of hidden Markov trees)**.** Let $\mathcal{T}$ be a rooted tree on the index set $[L]$, with root $o \in [L]$ and parent map $\mathrm{pa}(\cdot)$ for nonroot vertices (see Appendix A.1 for the formal definition). A J-state *hidden Markov tree* on $\mathcal{T}$ is a distribution q on $\mathbb{V}^L$ of the form

$$q(x) = \sum_{z\in[J]^L} q(x,z) \qquad \text{with} \qquad q(x,z) := \nu(z_o) \prod_{i\neq o} A_i\big(z_i, z_{\mathrm{pa}(i)}\big) \prod_{i=1}^{L} B_i(x_i, z_i). \tag{11}$$

Here, ν is the root-state distribution on $[J]$. For hidden state $b \in [J]$, the transition kernel $A_i(\cdot, b)$ is a distribution on $[J]$ for $i \neq o$, and the emission kernel $B_i(\cdot, b)$ is a distribution on $\mathbb{V}$ for $i \in [L]$.

The class $\mathcal{M}_{J,R}$ consists of all mixtures of R such models:

$$q(x) = \sum_{r=1}^{R} \pi_r q_r(x) \qquad \text{with} \qquad \pi_r \geq 0, \quad \sum_{r=1}^{R} \pi_r = 1,$$

where each q_r is a J-state hidden Markov tree with its own topology $\mathcal{T}_r$, root distribution, transition kernels, and observation kernels.

The MHMT has a simple generative interpretation. We first draw a component label $C \sim \pi$. Within that component, we draw $Z_o \sim \nu$, then, moving away from the root, $Z_i \sim A_i(\cdot, b)$ given that its parent is in state b, and finally each token independently as $X_i \sim B_i(\cdot, a)$ given $Z_i = a$. For the chain, $o = 1$, $\mathrm{pa}(i) = i - 1$, so a HMT only replaces the predecessor $i - 1$ by the parent $\mathrm{pa}(i)$, which need not be adjacent to i in the sequence. Therefore, the token X_i may not only depend on its immediate predecessor X_{i-1}, supporting long-range and bi-directional dependence patterns.

Remark 1 (Why mixture?). For fixed component topologies, a distribution in the class $\mathcal{M}_{J,R}$ has $O\big(RLJ(J+V)\big)$ parameters. This representation complexity can be much smaller than the support size: even a single component with strictly positive observation kernel has full support on $\mathbb{V}^L$. Consequently, guarantees that scale with the support size may incur a dependence on V^L, even when the distribution admits a compact representation with small J and R. Mixtures enrich this representation by allowing components to differ in tree topology, transitions, and emissions (Meila & Jordan, 2000), while increasing the parameter count only linearly in R. To illustrate this, suppose that all components share a common tree topology. The mixture can then be represented as a single HMT with RJ states with block diagonal transition matrices. This structure requires only $O(LRJ^2)$ transition parameters, reduced from $O(LR^2J^2)$ for a generic RJ-state HMT with dense transitions.

Approximation of low-dimensional data. We focus on the distribution $p^\star$ that is close to the MHMT class $\mathcal{M}_{J,R}$, allowing departures from exact tree-mixture structure. Misspecification enters through the following approximation error.

Definition 2 (KL approximation error)**.** Let $\mathcal{F}$ be a class of probability distributions on $\mathbb{V}^L$. For any distribution p on $\mathbb{V}^L$, define the KL approximation error w.r.t. $\mathcal{F}$ by

$$\mathsf{Proj}_{\mathcal{F}}(p) := \inf_{q\in\mathcal{F}} D_{\mathrm{KL}}\big(p\|q\big). \tag{12}$$

In particular, we write $\mathsf{Proj}_{J,R} := \mathsf{Proj}_{\mathcal{M}_{J,R}}$ for the MHMT class $\mathcal{M}_{J,R}$.

The quantity $\mathsf{Proj}_{J,R}(p^\star)$ measures how well $p^\star$ can be approximated by a R-mixture J-state hidden markov tree models. The pair (J,R) indexes the approximation class used in our analysis and is not an input to the training algorithm once the Transformer architecture is fixed. Increasing J or R can reduce the approximation error at the cost of greater model complexity.

4 Main results: statistical guarantees for MDMs

In this section, we first develop estimation error bounds for the Transformer mask predictor learned by ERM, and then establish end-to-end sampling guarantees for the resulting MDM-based sampler.

4.1 Learning mask predictor

Fix integers $n, L, V \geq 2$ and $J, R \geq 1$. Let $s = (s_1, \cdots, s_T)$ be a fixed-cardinality unmasking schedule with $s_{\mathsf{max}} := \max s_t$. We use the same schedule s for training and inference.

Recall from Section 2 that the ERM $\widehat{\theta}$ minimizes the empirical masked-token negative log-likelihood, that the excess risk $\mathcal{L}(\theta) := \overline{R}(p_\theta) - \overline{R}(p^\star)$, with $\overline{R}(q)$ defined in (3),equals the average KL divergence between the true and predicted conditionals of masked tokens.

The following theorem controls this error of the Transformer mask predictor learned by ERM. The proof can be found in Appendix A.4 and is sketched in Section 5.

Theorem 1 (Estimation error of mask predictor)**.** *There exists a transformer class Θ of the form* (9)*, with an architecture depending only on (n, L, V, J, R) (specified in Appendix C.2), such that for any distribution $p^\star$ on $\mathbb{V}^L$, the ERM $\widehat{\theta}$ over Θ computed from n i.i.d. samples drawn from $p^\star$ satisfies*

$$\mathbb{E}[\mathcal{L}(\widehat{\theta})] \leq C_1\left(\frac{s_{\mathsf{max}}}{L}\mathsf{Proj}_{J,R}(p^\star) + \frac{K_{\mathrm{par}}}{n}\right)\mathsf{poly}\log(n, L, V, J, R). \tag{13}$$

Here, $C_1 > 0$ is a universal constant, the expectation is over training samples, and

$$K_{\mathrm{par}} := R[LJ(J+V) + J^3]. \tag{14}$$

The two terms in the error bound (13) have distinct roles. The first term captures the approximation error of $p^\star$ by the MHMT class $\mathcal{M}_{J,R}$, and the factor s_{max}/L converts the joint KL error into a per-token one. The second term corresponds to the statistical error due to finite-sample training. The network has polylogarithmic attention blocks, at most $2R$ attention heads, and $\widetilde{O}(K_{\mathrm{par}})$ nonzero parameters, so the statistical term is a parameter count over the sample size n.

When the number of hidden states is moderate in the sense that $J \lesssim \sqrt{LV}$, where L is the sequence length and V is the vocabulary size, $K_{\mathrm{par}} = O\big(RLJ(J+V)\big)$ of the same order as the number of parameters of $\mathcal{M}_{J,R}$ (see Section 3). Up to logarithmic factors, the statistical error is thus governed by this intrinsic complexity rather than by the ambient dimension $V^L - 1$. Since the architecture depends only on (n, L, V, J, R), this holds without knowledge of the trees, parameters, or latent states of the approximating mixture. Transformer adapts to the unknown latent structure.

4.2 End-to-end sampling guarantees

Recall from Section 2 a T-step MDM sampler with a fixed-cardinality unmasking schedule $s = (s_1, \ldots, s_T)$. It outputs a sequence Y in $\mathbb{V}^L$ by iteratively unmasking tokens in the index set S^t in parallel according to a mask predictor p_θ at step t, where S^t with size $|S^t| = s_t$ is a random subset of the currently masked positions.

To isolate the effect of parallel sampling, let $Y^\star$ denote the output of the same MDM sampler when the exact conditional marginals of $p^\star$ is used as the mask predictor. Define the parallelization error

$$\varepsilon_{\mathsf{parall}}(p^\star; s) := \mathbb{E}_S[D_{\mathrm{KL}}(p^\star \,\|\, p_{Y^\star|S})], \tag{15}$$

where the expectation is over the random unmasking sets $S = (S^1, \ldots, S^T)$ with $|S^t| = s_t$.

We now present the end-to-end sampling guarantee for the MDM-based sampling. The proof can be found in Appendix A.5.

Theorem 2 (End-to-end sampling error of MDM)**.** *Instate the conditions and notation in Theorem 1. Let $p_{Y|S}$ denote the output distribution of the T-step MDM sampler using the ERM-based mask predictor $p_{\widehat{\theta}}$, the realized unmasking sets $\mathcal{S} = (S^1, \ldots, S^T)$ with $|S^t| = s_t$. Then for any $1 \le T \le L$, one has*

$$\begin{aligned}&\mathbb{E}_{\{X^{(i)}\}_{i=1}^n, S}\big[D_{\mathrm{KL}}\big(p^\star \,\|\, p_{Y|S}\big)\big] \\ &\quad\le C_2\left(\varepsilon_{\mathsf{parall}}(p^\star; s) + s_{\mathsf{max}}\mathsf{Proj}_{J,R}(p^\star) + \frac{LK_{\mathrm{par}}}{n}\right)\mathsf{poly}\log(n, L, V, J, R),\end{aligned} \tag{16}$$

where $C_2 > 0$ is a universal constant, and the expectation is over the training samples and random unmasking sets S.

To the best of our knowledge, this is the first end-to-end guarantee for MDMs that consider both factorization error due to parallel sampling and finite-sample estimation error of the mask predictor. Moreover, the guarantee is adaptive to the low-dimensional structure of the data distribution $p^\star$.

The last two terms in the bound (16) are inherited from the estimation error of the mask predictor in Theorem 1, multiplied by L to convert the per-token KL error into a joint KL error. They capture the impact of finite-sample mask predictor training on the end-to-end sampling error. The first term $\varepsilon_{\mathsf{parall}}(p^\star; s)$ captures the parallelization error due to simultaneous sampling of multiple tokens that ignore their dependencies, which depends only on the data distribution $p^\star$ and the unmasking schedule s used in the inference stage.

Remark 2 (Parallelization error). Recent sampling analysis (Li & Cai, 2025; Wainwright, 2026) characterizes the parallelization error $\varepsilon_{\mathsf{parall}}(p^\star; s)$ in terms of the information-theoretic structure of the data distribution $p^\star$ and the unmasking schedule s. For instance, as shown in Chen et al. (2025), for any $T = \Omega(\log L)$, there exist a schedule s^{DTC} [2] depending only on (L, T) such that

$$\varepsilon_{\mathsf{parall}}(p^\star; s^{\mathsf{DTC}}) \le \frac{6\log(eL)}{T}\,\mathsf{DTC}(p^\star). \tag{17}$$

Here *dual total correlation* (DTC) is defined as $\mathrm{DTC}(p^\star) := H(X) - \sum_{i=1}^L H(X_i \mid X_{[L]\setminus\{i\}})$, where $X \sim p^\star$ and H denotes the entropy. In particular, DTC can be substantially smaller than the sequence length L for structured data distributions, e.g., a MHMT with star-shaped components satisfies $\mathrm{DTC}(p^\star) \le \log(RJ)$. We refer to Chen et al. (2025); Dmitriev et al. (2026); Zhao & Cai (2026) for further discussion of distributions with small TC and DTC.

We now discuss several important implications of Theorem 2.

[2] The original construction in Chen et al. (2025) requires an estimate of DTC. A straightforward modification yields a schedule that depends only on (L, T) with the same guarantees. For completeness, we present the construction s^{DTC} in Appendix A.3.

Fast and data-efficient sampling for structured data. Assume that $p^\star \in \mathcal{M}_{J,R}$, i.e., the data distribution is exactly a R-mixture of J-state hidden Markov trees. Use s^{DTC} from Remark 2 with T steps. Then Theorem 2 yields

$$\mathbb{E}_{\{X^{(i)}\}_{i=1}^n,S}\big[D_{\mathrm{KL}}\big(p^\star \,\|\, p_{Y|S}\big)\big] \lesssim \left(\frac{\mathrm{DTC}(p^\star)}{T} + \frac{LK_{\mathrm{par}}}{n}\right)\mathsf{poly}\log(n, L, V, J, R). \tag{18}$$

As a result, when the latent space is small compared to the vocabulary size and length, i.e., $J \lesssim \min\{V, \sqrt{LV}\}$, to generate a sample that is ε-close to $p^\star$ in KL divergence, it suffices to take

$$\text{number of iterations: } T = \widetilde{O}\Big(\frac{\mathrm{DTC}(p^\star)}{\varepsilon}\Big), \quad \text{number of samples: } n = \widetilde{O}\Big(\frac{L^2RJV}{\varepsilon}\Big). \tag{19}$$

This shows that the MDM can achieve sampling acceleration and avoid the curse of dimensionality in sample complexity when the data distribution has low-dimensional latent structure. For instance, when $p^\star$ is a mixture of R star-shaped HMMTs with J hidden states, $\mathrm{DTC}(p^\star) \le \log(RJ)$, and the number of iterations scales logarithmically in (R, J), up to logarithmic dependence on the sequence length L. The number of training samples polynomially scales with (L, V, J, R). This shows that the MDM can efficiently learn and sample from structured data distributions.

Adaptation to low-complexity structure. The architecture in Theorems 1–2 depends only on (n, L, V, J, R), not on $p^\star$ or on the latent variables, trees, and parameters, none of which are observed. One network class approximates the masked-token conditionals of every $q \in \mathcal{M}_{J,R}$ uniformly. In this sense, transformer-based MDMs adapt to unknown latent structure, with (J, R) acting as a budget parameter that trades approximation against estimation.

Handling non-exact HMMT structure. By Theorem 2, for general data distributions $p^\star$ that do not exactly belong to the MHMT class $\mathcal{M}_{J,R}$, to achieve ε-sampling accuracy, the sample complexity polynomially scales with the parameter (J, R) that ensure $\mathsf{Proj}_{J,R}(p^\star) \lesssim \varepsilon/s_{\mathsf{max}}$.

We show two examples show that the required (J, R) can be chosen to scale with low-dimensional structure of the data distribution, rather than the V^L. The detailed discussions and proofs are deferred to Appendix A.2.

Continuous latent representation. Suppose that the data distribution satisfies $p^\star(x) = \int_{[0,1]^k} \prod_{i=1}^L p_i(x_i \mid z)\mu(\mathrm{d}z)$ for a latent distribution μ on $[0,1]^k$, where the kernels $p_i(\cdot \mid z)$ are smooth in z in the sense that for some $\alpha \in (0, 1]$, $D_{\mathrm{KL}}\big(p_i(\cdot \mid z) \,\|\, p_i(\cdot \mid z')\big) \lesssim \|z - z'\|_2^{2\alpha}$ for $i \in [L]$ and $z, z' \in [0,1]^k$ This covers quantized Lipschitz decoders with independent Gaussian observation noise Dmitriev et al. (2026). Discretizing μ onto a grid of size R gives a mixture of R product distributions with $\mathsf{Proj}_{J,R}(p^\star) \lesssim LR^{-2\alpha/k}$ for any $J \ge 1$, so

$$R \gtrsim (\varepsilon^{-1}s_{\max}L)^{k/(2\alpha)} \Longrightarrow \mathsf{Proj}_{J,R}(p^\star) \le \varepsilon/s_{\max}.$$

So the component count is governed by the latent dimension k and independent of V.

Low Rényi entropy. For $\alpha \in (0, 1)$, the Rényi entropy $H_\alpha(p) := \frac{1}{1-\alpha}\log\sum_{x\in\mathbb{V}^L} p^\alpha(x)$ is small when p concentrates its mass on few points. In particular, $H_\alpha(p) \uparrow \log|\mathsf{supp}(p)|$ as $\alpha \to 0$. For any integer $R \ge 2$ and $J \ge 1$, one can show that $\mathsf{Proj}_{J,R}(p^\star) \le L\log V\big(e^{H_\alpha(p^\star)}/R\big)^{(1-\alpha)/\alpha}$, that is,

$$R \ge e^{H_\alpha(p^\star)}(\varepsilon^{-1}s_{\max}L\log V)^{\alpha/(1-\alpha)} \Longrightarrow \mathsf{Proj}_{J,R}(p^\star) \le \varepsilon/s_{\max}.$$

So the component count is controlled by the Rényi entropy and $L\log V$.

5 Analysis

In this section, we outline the proof of Theorem 1. Recall the definition of $\mathcal{L}(\theta) = \overline{R}(p_\theta) - \overline{R}(p^\star)$, where $\overline{R}(q) := \mathbb{E}_{X\sim p^\star}[\ell_q(X)]$ and ℓ_q is defined in (3). The estimation error can be decomposed as

$$\mathcal{L}(\widehat{\theta}) = \underbrace{\overline{R}(p_{\widehat{\theta}}) - \inf_{\theta\in\Theta}\overline{R}(p_\theta)}_{\text{generalization error}} + \underbrace{\inf_{\theta\in\Theta}\overline{R}(p_\theta) - \overline{R}(p^\star)}_{\text{approximation error: } \varepsilon_{\mathsf{approx}}(\Theta)}. \tag{20}$$

We bound these two terms as follows, respectively.

Generalization error. By the definition of ERM estimator in Section 2, the generalization error can be controlled by,

$$\big|\overline{R}(p_{\widehat{\theta}}) - \inf_{\theta\in\Theta} \overline{R}(p_\theta)\big| \lesssim \sup_{\theta\in\Theta} \big|\widehat{R}_n(p_\theta) - \overline{R}(p_\theta)\big|,$$

with the empirical risk $\widehat{R}_n$ defined in Section 2. Thus, the generalization error is governed by the properties of the loss and the capacity of the function class. Using the localization techniques of Bartlett et al. (2005); Koltchinskii (2006), we obtain the following bound.

Proposition 1 (Localized oracle inequality for a logit class)**.** *Fix integers $n, V \geq 3$, and a unmasking schedule s in Section 2.1. Let $\widehat{\theta}$ be an exact ERM as in Section 2. Define $M_G := G + \log V$ and suppose Assumption 1 holds. Due to space constraints, the definitions of G and $\mathcal{C}(\Theta)$, together with the statement of Assumption 1, are deferred to Appendix B.2. Then*

$$\mathbb{E}\big[\mathcal{L}(\widehat{\theta})\big] \lesssim \varepsilon_{\mathsf{approx}}(\Theta) + M_G \cdot \frac{C(\Theta)\log(en)+1}{n}. \tag{21}$$

Here the expectation is taken over training data $\{X^{(j)}\}_{j=1}^n$.

We further prove the capacity bound $C(\Theta)$ for our sparse transformer class in Lemma 5.

Approximation error. We next control the approximation error by constructing sparse transformer-based predictors. The architectural budgets and proof of the following proposition are given in Appendix C.2.

Proposition 2 (Tree-mixture approximation)**.** *Fix integers $L, V \geq 2$, $J, R \geq 1$, and $0 < \delta, \epsilon \leq 1/2$. There exists a sparse Transformer class Θ, with architectural budgets depending only on $(L, V, J, R, \delta, \epsilon)$, such that*

$$\varepsilon_{\mathsf{approx}}(\Theta) \leq \frac{s_{\max}}{L}\mathsf{Proj}_{J,R}(p^\star) + s_{\max}\log\frac{1}{1-\delta} + 2\epsilon \tag{22}$$

for every distribution $p^\star$ on $\mathbb{V}^L$.

To obtain this bound, we first regularize any intermediate distribution $q \in \mathcal{M}_{J,R}$ by independently replacing each token with a uniform token with probability δ, as in Appendix B.4. This procedures preserves the tree-mixture structure. Then we control the error brought by regularization and further decompose the approximation error in Appendix B.5.

For a single tree, we construct the transformer-based predictor with depth $\widetilde{O}(\log L)$ by contracting tree vertices in parallel and then reversing these operations. Appendix C.2 extends the construction to mixtures by combining the component conditionals with weights determined by context likelihoods. For shallow-tree cases, Appendix C.5 provides an alternative construction whose depth depends on the tree height.

6 Discussion

This paper studies the sample complexity of masked diffusion models (MDMs) for learning distributions with low-complexity structure, measured by their approximation through mixtures of hidden markov tree models. We establish end-to-end sampling guarantees whose statistical error depends on the structural complexity $\mathrm{poly}(L, V, J, R)$, rather than the size of sequence space V^L. Using the prescribed architectural parameters, our transformer predictor adapts to the unknown latent structure through the training objective on observed data. These results helps explain why transformer-based MDMs performs well in learning distributions with structural dependence.

Several directions remain open. First, extending the analysis beyond mixtures of hidden markov trees would accommodate more general dependence structures. Second, establishing statistical guarantees for confidence-based decoding would clarify how adaptive token selection guided by the learned predictor affects the final sampling error. Third, incorporating optimization error would further help bridge the gap between theoretical guarantees with practical training.

AI USE STATEMENT

In this work, we used generative AI tools for [formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, interpret results]. We have not used generative AI tools for [generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis]. Additionally, we used generative AI tools for [literature review, proof check]. We have reviewed all AI-assisted work. We have checked the correctness of all LLM-polished proofs and have done further literature review. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*, 2024.

Yoseph Barash, Gal Elidan, Nir Friedman, and Tommy Kaplan. Modeling dependencies in protein-dna binding sites. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, pp. 28–37, 2003.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. 2005.

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

Changxiao Cai and Gen Li. Minimax optimality of the probability flow ode for diffusion models. *arXiv preprint arXiv:2503.09583*, 2025.

Changxiao Cai and Gen Li. Confidence-based decoding is provably efficient for diffusion language models. *arXiv preprint arXiv:2603.22248*, 2026.

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

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

Sitan Chen, Kevin Cong, and Jerry Li. Optimal inference schedules for masked diffusion models. *arXiv preprint arXiv:2511.04647*, 2025.

Justin Chiu and Alexander M Rush. Scaling hidden markov language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1341–1349, 2020.

Cholyeon Cho and Yuchen Wu. Minimax optimality of score-entropy discrete diffusion. *arXiv preprint arXiv:2608.20635*, 2026.

Matthew S Crouse, Robert D Nowak, and Richard G Baraniuk. Wavelet-based statistical signal processing using hidden markov models. *IEEE Transactions on signal processing*, 46(4):886–902, 1998.

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

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

Shansan Gong, Ruixiang Zhang, Huangjie Zheng, Jiatao Gu, Navdeep Jaitly, Lingpeng Kong, and Yizhe Zhang. Diffucoder: Understanding and improving masked diffusion models for code generation. In *International Conference on Learning Representations*, volume 2026, pp. 15723–15750, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. 2006.

Gen Li and Changxiao Cai. Breaking ar's sampling bottleneck: Provable acceleration via diffusion language models. *Advances in Neural Information Processing Systems*, 38:11700–11725, 2025.

Yang Li, Shaoyang Ning, Sarah E Calvo, Vamsi K Mootha, and Jun S Liu. Bayesian hidden markov tree models for clustering genes with shared evolutionary history. 2019.

Yuchen Liang, Zhiheng Tan, Ness Shroff, and Yingbin Liang. Sharp convergence rates for masked diffusion models. *arXiv preprint arXiv:2602.22505*, 2026.

Marina Meila and Michael I Jordan. Learning with mixtures of trees. *Journal of Machine Learning Research*, 1(Oct):1–48, 2000.

Gary L Miller and John H Reif. Parallel tree contraction and its application. In *FOCS*, volume 26, pp. 478–489, 1985.

Farnaz Mohammadi, Shakthi Visagan, Sean M Gross, Luka Karginov, JC Lagarde, Laura M Heiser, and Aaron S Meyer. A lineage tree-based hidden markov model quantifies cellular heterogeneity and plasticity. *Communications Biology*, 5(1):1258, 2022.

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. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pp. 26517–26582. PMLR, 2023.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *International Conference on Learning Representations*, volume 2025, pp. 64972–65009, 2025.

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

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Aadithya Srikanth, Mudit Gaur, and Vaneet Aggarwal. Discrete state diffusion models: A sample complexity perspective. *arXiv preprint arXiv:2510.10854*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Shintaro Wakasugi and Taiji Suzuki. State size independent statistical error bound for discrete diffusion models. *Advances in Neural Information Processing Systems*, 38:138908–138943, 2026.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. *arXiv preprint arXiv:2402.18567*, 2024.

Jingda Wu and Changxiao Cai. Neural network-based diffusion models adapt to low-dimensional multi-modal data structure. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*.

Jingda Wu and Changxiao Cai. Diffusion models are statistically optimal for learning low-dimensional multi-modal distributions. *arXiv preprint arXiv:2605.30153*, 2026.

Alexander Ypma and Tom Heskes. Automatic categorization of web pages and user clustering with mixtures of hidden markov models. In *International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles*, pp. 35–49. Springer, 2002.

Oussama Zekri, Ambroise Odonnat, Abdelhakim Benechehab, Linus Bleistein, Nicolas Boullé, and Ievgen Redko. Large language models as markov chains. *arXiv preprint arXiv:2410.02724*, 2024.

Zixuan Zhang, Hengyu Fu, Zhuoran Yang, Mengdi Wang, Tuo Zhao, and Minshuo Chen. Generalization bounds for discrete diffusion: Statistical advantage of masking. In *Forty-third International Conference on Machine Learning*.

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

A Examples and Proofs of theorems

A.1 Formal definition for tree model

Definition 3 (Dependence trees and tree distributions)**.** Let $\mathcal{T}$ be a rooted tree structure on the vertex set $[L]$, namely a connected, acyclic graph with a designated root $o \in [L]$. For each $i \in [L] \setminus \{o\}$, denote $\mathrm{pa}(i)$ as its unique parent, which represents the edge $(i, \mathrm{pa}(i))$. Then the tree structure $\mathcal{T}$ can be specified as

$$\mathcal{T} := \big\{\{o\}, \big(i, \mathrm{pa}(i)\big)_{i\in[L]\setminus\{o\}}\big\}. \tag{23}$$

In particular, repeatedly following the parent map from any nonroot vertex reaches o after finite steps. We further denote $\mathfrak{T}$ as the set of all trees on $[L]$ and suppress its dependence on L.

For any discrete random vector $Z = (Z_1, \cdots, Z_L)$, we say that Z follows a tree distribution if there exists a dependence tree $\mathcal{T} \in \mathfrak{T}$ such that the law of Z admits the following factorization

$$p_Z(z_1, \cdots, z_L) = p_o(z_o) \prod_{i\neq o} p_i(z_i \mid z_{\mathrm{pa}(i)}). \tag{24}$$

Here the root o and parent map $\mathrm{pa}(i)$ corresponds to tree structure $\mathcal{T}$. We further say that p_Z factories according to $\mathcal{T}$.

A.2 Details of examples in Section 4.2.

In this section, we derive the bounds on $\mathsf{Proj}_{J,R}(p^\star)$ for the continuous latent representation and low Rényi entropy examples in Section 4.2.

Continuous latent representation.

Lemma 1. *Suppose that there exists a probability distribution μ on $[0,1]^k$ such that,*

$$p^\star(x) = \int_{[0,1]^k} \prod_{i=1}^{L} p_i(x_i \mid z)\mu(\mathrm{d}z), \quad x \in \mathbb{V}^L \tag{25}$$

We further assume that for some $A \geq 0$ and $\alpha \in (0, 1]$, the kernel admits the following continuity condition:

$$D_{\mathrm{KL}}\big(p_i(\cdot \mid z) \,\|\, p_i(\cdot \mid z')\big) \leq A^2 \|z - z'\|_2^{2\alpha}, \quad \forall i \in [L], \quad z, z' \in [0,1]^k. \tag{26}$$

Then for any integer $J, R \geq 1$, it holds that,

$$\mathsf{Proj}_{J,R}(p^\star) \leq \mathsf{Proj}_{1,R}(p^\star) \leq C_{k,\alpha} L A^2 R^{-2\alpha/k}. \tag{27}$$

Here constant $C_{k,\alpha}$ only depends on k, α. In particular, condition (26) holds if $p_i(\cdot \mid z) = \mathrm{softmax}(g_i(z))$ with logits $g_i : [0,1]^k \to \mathbb{R}^V$ satisfying

$$\|g_i(z) - g_i(z')\|_2 \leq A\|z - z'\|_2^\alpha, \quad \forall i \in [L], \quad z, z' \in [0,1]^k \tag{28}$$

Proof of Lemma 1. We approximate the latent variable by a finite grid. Let $s = \lfloor R^{1/k} \rfloor \geq 1$, and partition $[0,1]^k$ into $s^k \leq R$ cubes $\{C_j\}_{j=1}^{s^k}$ of side length $1/s$, assigning their boundaries so that the cells are disjoint. Denote by z_j the midpoint of C_j, and define

$$w_j := \mu(C_j), \qquad q_R(x) := \sum_{j=1}^{s^k} w_j \prod_{i=1}^{L} p_i(x_i \mid z_j).$$

Since each component is a product distribution, $q_R \in \mathcal{M}_{1,R}$. For $z \in [0,1]^k$, let $j(z)$ be the index of its cell. Consider the joint distributions

$$P(\mathrm{d}z, x) := \mu(\mathrm{d}z) \prod_{i=1}^{L} p_i(x_i \mid z), \qquad Q(\mathrm{d}z, x) := \mu(\mathrm{d}z) \prod_{i=1}^{L} p_i(x_i \mid z_{j(z)}).$$

Their marginals on x are $p^\star$ and q_R, respectively. By the data-processing inequality and the chain rule for KL divergence,

$$\begin{aligned} D_{\mathrm{KL}}(p^\star \| q_R) &\leq D_{\mathrm{KL}}(P \| Q) \\ &= \int_{[0,1]^k} \sum_{i=1}^{L} D_{\mathrm{KL}}\big(p_i(\cdot \mid z) \,\|\, p_i(\cdot \mid z_{j(z)})\big)\, \mu(\mathrm{d}z) \\ &\leq LA^2 \int_{[0,1]^k} \|z - z_{j(z)}\|_2^{2\alpha}\, \mu(\mathrm{d}z) \\ &\leq LA^2 \left(\frac{\sqrt{k}}{2s}\right)^{2\alpha} \leq k^\alpha L A^2 R^{-2\alpha/k}. \end{aligned}$$

Here the last two inequalities use the cell diameter and $s \geq R^{1/k}/2$. Together with $\mathcal{M}_{1,R} \subseteq \mathcal{M}_{J,R}$, this proves (27) with $C_{k,\alpha} = k^\alpha$.

It remains to verify the sufficient condition (28). Let $F(a) := \log \sum_{v=1}^{V} e^{a_v}$. For $u = \mathrm{softmax}(a)$, we have

$$\nabla F(a) = u, \qquad \nabla^2 F(a) = \mathrm{diag}(u) - uu^\top, \qquad \|\nabla^2 F(a)\|_{\mathrm{op}} \leq 1.$$

Hence, for any $a, b \in \mathbb{R}^V$,

$$\begin{aligned} D_{\mathrm{KL}}\big(\mathrm{softmax}(a) \,\|\, \mathrm{softmax}(b)\big) &= F(b) - F(a) - \nabla F(a)^\top (b - a) \\ &\leq \frac{1}{2}\|b - a\|_2^2. \end{aligned}$$

Taking $a = g_i(z)$ and $b = g_i(z')$, and applying (28) gives

$$D_{\mathrm{KL}}\big(p_i(\cdot \mid z) \,\|\, p_i(\cdot \mid z')\big) \leq \frac{1}{2}\|g_i(z) - g_i(z')\|_2^2 \leq \frac{A^2}{2}\|z - z'\|_2^{2\alpha}.$$

This implies (26) and completes the proof.

Low Rényi entropy. For a distribution p on $\mathbb{V}^L$ and $\alpha \in (0,1)$, recall that its Rényi entropy is defined by

$$H_\alpha(p) := \frac{1}{1-\alpha} \log \sum_{x \in \mathbb{V}^L} p(x)^\alpha.$$

Lemma 2 (Approximation under low Rényi entropy)**.** *Let $p^\star$ be a distribution on $\mathbb{V}^L$ and $\alpha \in (0,1)$. For any integers $J \geq 1$ and $R \geq 2$, it holds that*

$$\mathsf{Proj}_{J,R}(p^\star) \leq \mathsf{Proj}_{1,R}(p^\star) \leq L \log V \min\left\{1, \left(\frac{e^{H_\alpha(p^\star)}}{R}\right)^{(1-\alpha)/\alpha}\right\}. \tag{29}$$

Proof of Lemma 2. We maintain the $R-1$ largest atoms and approximate the remaining mass by a uniform distribution. Enumerate the points of $\mathbb{V}^L$ as $x_{(1)}, \ldots, x_{(V^L)}$, with probabilities $p_{(j)} := p^\star(x_{(j)})$ in decreasing order. Set $p_{(j)} = 0$ for $j > V^L$, and denote

$$S_\alpha := \sum_{j \geq 1} p_{(j)}^\alpha = e^{(1-\alpha)H_\alpha(p^\star)}, \quad \tau_{R-1} := \sum_{j \geq R} p_{(j)}.$$

If $\tau_{R-1} = 0$, then $p^\star$ is a mixture of at most $R-1$ point masses. Each point mass is a product distribution, so $\mathsf{Proj}_{1,R}(p^\star) = 0$.

For $\tau_{R-1} > 0$, let u be the uniform distribution on $\mathbb{V}^L$, and define

$$q_R := \sum_{j=1}^{R-1} p_{(j)} \delta_{x_{(j)}} + \tau_{R-1} u, \qquad v := \frac{1}{\tau_{R-1}} \sum_{j=R}^{V^L} p_{(j)} \delta_{x_{(j)}}.$$

Then $q_R \in \mathcal{M}_{1,R}$ and $p^\star = \sum_{j=1}^{R-1} p_{(j)} \delta_{x_{(j)}} + \tau_{R-1} v$. By joint convexity of KL divergence,

$$\begin{aligned} \mathsf{Proj}_{1,R}(p^\star) &\leq D_{\mathrm{KL}}(p^\star \| q_R) \leq \tau_{R-1} D_{\mathrm{KL}}(v \| u) \\ &= \tau_{R-1}\big(L \log V - H(v)\big) \leq \tau_{R-1} L \log V. \end{aligned}$$

To bound the remaining mass, notice that

$$\tau_{R-1} = \sum_{j \geq R} p_{(j)}^\alpha p_{(j)}^{1-\alpha} \leq p_{(R)}^{1-\alpha} S_\alpha, \quad \text{and } R\, p_{(R)}^\alpha \leq \sum_{j=1}^{R} p_{(j)}^\alpha \leq S_\alpha.$$

Therefore,

$$\tau_{R-1} \leq \left(\frac{S_\alpha}{R}\right)^{(1-\alpha)/\alpha} S_\alpha = \left(\frac{e^{H_\alpha(p^\star)}}{R}\right)^{(1-\alpha)/\alpha}.$$

Combining this bound with $\tau_{R-1} \leq 1$ and $\mathcal{M}_{1,R} \subseteq \mathcal{M}_{J,R}$ proves (29).

A.3 Derivation of the parallelization error bound (17)

In this section, we prove (17). We first write the parallelization error using conditional total correlation, and then construct the TC and DTC schedules following Section 5 of Chen et al. (2025).

Fix a distribution p on $\mathbb{V}^L$ and a forward masking schedule $\boldsymbol{s} = (s_1, \ldots, s_T)$, with $s_t \in \mathbb{N}_+$ and $\sum_{t=1}^T s_t = L$. Recall the index sets S^t and W^t defined in the reverse process in Section 2. The partition sets $\mathcal{S} := (S^1, \ldots, S^T)$ are selected uniformly with $|S^t| = s_t$, and

$$W^t = \bigcup_{u=t+1}^{T} S^u.$$

Thus, W^t contains the positions already generated before sampling S^t. Conditioned on $\mathcal{S}$, the sampler using the exact conditional marginals has output distribution

$$q_p^{\mathcal{S}}(x) := \prod_{t=1}^{T} \prod_{i \in S^t} p(x_i \mid x_{W^t}).$$

The parallelization error is

$$\varepsilon_{\mathrm{parallel}}(p;\boldsymbol{s}) := \mathbb{E}_{\mathcal{S}}\big[D_{\mathrm{KL}}(p\|q_p^{\mathcal{S}})\big] = \mathbb{E}_{\mathcal{S}}\Big[\sum_{t=1}^{T}\mathrm{TC}_p(X_{S_t}\mid X_{W^t})\Big], \tag{30}$$

where $X\sim p$ and

$$\mathrm{TC}_p(X_B\mid X_W) := \sum_{i\in B} H(X_i\mid X_W) - H(X_B\mid X_W).$$

The equality follows from the KL chain rule in reverse generation order.

Lemma 3 (TC and DTC schedules)**.** *Fix $L\geq 2$ and an integer T satisfying $2+2\log(eL)\leq T< L$. There exist forward masking schedules $\boldsymbol{s}^{\mathrm{TC}}$ and $\boldsymbol{s}^{\mathrm{DTC}}$, each consisting of T positive integers summing to L and depending only on (L,T), such that, for every distribution p on $\mathbb{V}^L$,*

$$\varepsilon_{\mathrm{parallel}}(p;\boldsymbol{s}^{\mathrm{TC}})\leq\frac{6\log(eL)}{T}\mathrm{TC}(p),\quad \varepsilon_{\mathrm{parallel}}(p;\boldsymbol{s}^{\mathrm{DTC}})\leq\frac{6\log(eL)}{T}\mathrm{DTC}(p). \tag{31}$$

For $T=L$, the singleton schedule has zero parallelization error.

Proof of Lemma 3. We first write the parallelization error in terms of average mutual information. For $j\in[L]$, define

$$Z_j := \mathbb{E}_{\substack{|W|=j-1\\ i\notin W}} I(X_i;X_W),\qquad d_r := Z_{r+1}-Z_r,\quad r\in[L-1],$$

where (W,i) is uniform among pairs of the indicated sizes. For a fixed i, coupling nested uniform subsets of $[L]\setminus\{i\}$ gives $I(X_i;X_W)\leq I(X_i;X_{W'})$ whenever $W\subseteq W'$. Averaging over i therefore gives $0=Z_1\leq\cdots\leq Z_L$, so $d_r\geq 0$.

Let H_j be the average of $H(X_W)$ over uniform subsets $W\subseteq[L]$ of size j, with $H_0=0$. By the entropy chain rule,

$$Z_j = H_1+H_{j-1}-H_j,\qquad \sum_{j=1}^{L} Z_j = LH_1-H_L=\mathrm{TC}(p).$$

Moreover, $LZ_L=\sum_{i=1}^{L} I(X_i;X_{[L]\setminus\{i\}})=\mathrm{TC}(p)+\mathrm{DTC}(p)$. Hence,

$$\mathrm{TC}(p)=\sum_{r=1}^{L-1}(L-r)d_r,\qquad \mathrm{DTC}(p)=\sum_{r=1}^{L-1} r\,d_r. \tag{32}$$

For the construction below, we list batch sizes in generation order as $\boldsymbol{b}=(b_1,\dots,b_m)$, and write $\boldsymbol{b}^{\mathrm{rev}}=(b_m,\dots,b_1)$ for the corresponding forward schedule. Set $N_j:=\sum_{\ell=1}^{j} b_\ell$ and $N_0=0$. For a batch with endpoints $a=N_{j-1}$ and $b=N_j$, the expected sum of conditional marginal entropies is $(b-a)(H_1-Z_{a+1})$, while the expected joint conditional entropy is H_b-H_a. Thus its contribution to (30) is

$$\begin{aligned} E_{a,b} &:= (b-a)(H_1-Z_{a+1})-(H_b-H_a)\\ &= \sum_{v=a+1}^{b}(Z_v-Z_{a+1}) = \sum_{r=a+1}^{b-1}(b-r)d_r. \end{aligned} \tag{33}$$

Summing over batches gives

$$\varepsilon_{\mathrm{parallel}}(p;\boldsymbol{b}^{\mathrm{rev}})=\sum_{j=1}^{m} E_{N_{j-1},N_j}. \tag{34}$$

In particular, singleton batches contribute zero.

TC schedule. Fix an integer $2 \le \zeta \le L$. Starting with $r_0 = L$ remaining positions, take a batch of size $\lfloor r_j/\zeta \rfloor$ whenever $r_j \ge \zeta$, so that

$$r_{j+1} = r_j - \lfloor r_j/\zeta \rfloor.$$

Once fewer than ζ positions remain, generate them in singleton batches. Denote this list of sizes by $\boldsymbol{b}^{\mathrm{TC}}$. Since r_j and ζ are integers,

$$r_{j+1} - \zeta + 1 \le \left(1 - \frac{1}{\zeta}\right)(r_j - \zeta + 1).$$

Therefore, at most

$$\lambda := \left\lfloor \frac{\log(L - \zeta + 1)}{-\log(1 - 1/\zeta)} \right\rfloor + 1 \le 1 + \zeta \log L$$

recursive steps are needed to reach $r_j < \zeta$. Including the remaining singleton batches, the total number of batches satisfies $m \le \lambda + \zeta - 1 \le \zeta \log(eL)$.

For a nonsingleton batch with endpoints a, b, its size satisfies

$$b - a \le \frac{L - a}{\zeta}, \qquad \text{which implies} \qquad b - a \le \frac{L - b}{\zeta - 1}.$$

Thus $b - r \le (L - r)/(\zeta - 1)$ for every $a < r < b$. Since these index ranges are disjoint across batches and $d_r \ge 0$, (32)–(34) give

$$\varepsilon_{\mathrm{parallel}}\big(p; (\boldsymbol{b}^{\mathrm{TC}})^{\mathrm{rev}}\big) \le \frac{1}{\zeta - 1} \sum_{r=1}^{L-1} (L - r) d_r = \frac{\mathrm{TC}(p)}{\zeta - 1}. \tag{35}$$

DTC schedule. Take $\boldsymbol{b}^{\mathrm{DTC}} = (\boldsymbol{b}^{\mathrm{TC}})^{\mathrm{rev}}$, which has the same number of batches. A nonsingleton batch with endpoints a', b' in the original list now has endpoints $a = L - b'$ and $b = L - a'$. The preceding size bound gives

$$b - a = b' - a' \le \frac{L - b'}{\zeta - 1} = \frac{a}{\zeta - 1}.$$

Hence $b - r \le r/(\zeta - 1)$ for $a < r < b$. Applying (32)–(34) again yields

$$\varepsilon_{\mathrm{parallel}}\big(p; (\boldsymbol{b}^{\mathrm{DTC}})^{\mathrm{rev}}\big) \le \frac{1}{\zeta - 1} \sum_{r=1}^{L-1} r\, d_r = \frac{\mathrm{DTC}(p)}{\zeta - 1}. \tag{36}$$

Finally, we choose ζ and refine the two lists to obtain exactly T batches. Let

$$A_L := \log(eL), \qquad \zeta := \left\lfloor \frac{T - 2}{A_L} \right\rfloor.$$

The assumption on T ensures $2 \le \zeta \le L$, and each list has at most $\zeta A_L \le T - 2$ batches. Notice that splitting a batch at an integer $a < c < b$ gives

$$E_{a,b} - E_{a,c} - E_{c,b} = (b - c)(Z_{c+1} - Z_{a+1}) \ge 0.$$

Thus, splitting cannot increase the parallelization error. Since $T < L$, we may repeatedly split the first nonsingleton batch into a singleton and its remainder until each list has exactly T batches. Denote the refined lists by $\widetilde{\boldsymbol{b}}^{\mathrm{TC}}$ and $\widetilde{\boldsymbol{b}}^{\mathrm{DTC}}$, and set

$$\boldsymbol{s}^{\mathrm{TC}} := (\widetilde{\boldsymbol{b}}^{\mathrm{TC}})^{\mathrm{rev}}, \qquad \boldsymbol{s}^{\mathrm{DTC}} := (\widetilde{\boldsymbol{b}}^{\mathrm{DTC}})^{\mathrm{rev}}.$$

Both forward schedules depend only on (L, T). For $x = (T - 2)/A_L \ge 2$, the inequality $\lfloor x \rfloor - 1 \ge x/3$ implies

$$\frac{1}{\zeta - 1} \le \frac{3A_L}{T - 2} \le \frac{6A_L}{T},$$

where the last inequality uses $T \ge 4$. Substituting this into (35) and (36) proves (31). For $T = L$, take singleton batches, for which every term in (34) is zero.

A.4 Proof of Theorem 1

In this section, we combine the approximation construction in Appendix C.2 with the capacity bound and the oracle inequality. We then give the corresponding bound for trees of height at most h.

Fix the deterministic masking schedule $\boldsymbol{s}$ and write

$$\delta = \varepsilon = \frac{1}{n}, \qquad G_n := \log(nV), \qquad \beta := \log(enLVJR), \qquad K_{\mathrm{par}} := R\big[L(J^2 + JV) + J^3\big].$$

Take the common parameter class Θ from Proposition 3 with these choices. Its budgets depend only on (n, L, V, J, R), and $G_\delta = G_n$. Moreover,

$$H_R = (L+1)\log(nV) + \log(64LJRn) \le CL\beta, \qquad b_R = \log(eLJVRH_Rn) \le C\beta.$$

The parameter class is a finite union of closed bounded sets, one for each allowed support, and is therefore compact. The empirical loss is continuous in the parameters, so an exact ERM $\widehat{\theta} \in \Theta$ exists.

Approximation error. The same class approximates $F^{\delta,q}$ uniformly over $q \in \mathcal{M}_{J,R}$. Lemma 7 therefore gives

$$\begin{aligned}\varepsilon_{\mathrm{approx}}(\Theta) &\le \frac{s_{\max}}{L}\mathsf{Proj}_{J,R}(p^\star) + s_{\max}\log\frac{1}{1-1/n} + \frac{2}{n} \\ &\le \frac{s_{\max}}{L}\mathsf{Proj}_{J,R}(p^\star) + \frac{2s_{\max}+2}{n}. \end{aligned} \tag{37}$$

Here we use $-\log(1-x) \le 2x$ for $0 \le x \le 1/2$. The projection infimum need not be attained.

Capacity and estimation error. The architectural bounds in Proposition 3 imply

$$N \le C\beta^3, \qquad M \le 2R, \qquad S \le CK_{\mathrm{par}}\beta^{10}, \qquad \log B \le C\beta^2.$$

The additional JV term in its sparsity bound is absorbed by LJV. The width bounds also give $\log P_{\mathrm{tr}} \le C\beta$, where

$$P_{\mathrm{tr}} := (V+1)D + LD + N(3MD^2 + 2DD_{\mathrm{ff}}) + DV$$

counts all parameters, including the input and output embeddings. By Lemma 5, with $R_{\mathrm{embed}} = \sqrt{3}$,

$$\log A_{\mathsf{TF}} \le CN^2 \log\big(eNM(B \vee 1)(R_{\mathrm{embed}} \vee 1)\big) \le C\beta^8.$$

Writing $\bar{S} := \min\{S, P_{\mathrm{tr}}\}$, we may consequently take

$$C(\Theta) = \bar{S}\left[\log\frac{eP_{\mathrm{tr}}}{\bar{S}} + \log A_{\mathsf{TF}}\right] \le CK_{\mathrm{par}}\beta^{18}. \tag{38}$$

Together with clipping to $[-G_n, 0]$, this verifies Assumption 1 for the logit class on masked sequences.

Since $M_{G_n} = G_n + \log V = \log(nV^2) \le C\beta$, Proposition 1 yields

$$\begin{aligned}\mathbb{E}\big[\mathcal{L}(\widehat{\theta})\big] &\le 4\varepsilon_{\mathrm{approx}}(\Theta) + CM_{G_n}\frac{C(\Theta)\log(en)+1}{n} \\ &\le C\left\{\frac{s_{\max}}{L}\mathsf{Proj}_{J,R}(p^\star) + \frac{K_{\mathrm{par}}\beta^{20}}{n}\right\}.\end{aligned}$$

The expectation is over the independent training sequences. In the last inequality, we use (37)–(38), $\log(en) \le \beta$, and $s_{\max} \le L \le K_{\mathrm{par}}$ to absorb the remaining terms. Although Proposition 1 is stated for $n, V \ge 3$, its proof applies here for $n, V \ge 2$: the empirical-process bound requires only $n \ge 2$, and $M_{G_n} = \log(nV^2) > 1$. Suppressing the logarithmic factor proves Theorem 1.

Trees of bounded height. For $1 \le h \le L-1$, use the common class Θ_h from Proposition 4, again with $\delta = \varepsilon = 1/n$. Lemma 7 gives

$$\varepsilon_{\text{approx}}(\Theta_h) \le \frac{s_{\max}}{L} \mathsf{Proj}_{\mathcal{M}^h_{J,R}}(p^\star) + \frac{2s_{\max}+2}{n}.$$

The height-dependent budgets satisfy

$$N \le C(h+1)\beta^2, \qquad S \le CRL(J^2+JV)\beta^8, \qquad \log B \le C\beta^2, \qquad \log P_{\text{tr}} \le C\beta.$$

Thus the same covering argument gives

$$C(\Theta_h) \le CRL(J^2+JV)(h+1)^2\beta^{14} \le CK_{\text{sh}}\beta^{14}, \qquad K_{\text{sh}} := RL(J^2+JV)h^2,$$

where $(h+1)^2 \le 4h^2$ because $h \ge 1$. Applying Proposition 1 to an exact ERM $\widehat{\theta}_h$ over Θ_h yields

$$\mathbb{E}\big[\mathcal{L}(\widehat{\theta}_h)\big] \le C\left\{ \frac{s_{\max}}{L} \mathsf{Proj}_{\mathcal{M}^h_{J,R}}(p^\star) + \frac{K_{\text{sh}}\beta^{16}}{n} \right\}.$$

This establishes the corresponding estimation bound for trees of bounded height.

A.5 Proof of Theorem 2

In this section, we decompose the sampling error into parallelization error and conditional prediction error, and then apply Theorem 1.

Fix the training sample and write $\theta = \widehat{\theta}$. Recall the ordered partition $\mathcal{S} = (S^1, \dots, S^T)$ and the unmasked index sets W^t in Appendix A.3. For the fixed schedule $\boldsymbol{s} = (s_1, \dots, s_T)$, each partition with $|S^t| = s_t$ has probability $\prod_{t=1}^T s_t!/L!$ under both forward masking and reverse generation. Thus, the two processes have the same distribution of index sets.

Conditioned on $\mathcal{S}$, denote the output distribution by

$$p_\theta^{\mathcal{S}}(x) := \prod_{t=1}^T \prod_{i \in S^t} p_{\theta,i}(x_i \mid x_{W^t}).$$

For a fixed context set $W \subseteq [L]$ and $i \notin W$, write

$$\Delta_{\theta,i}(W) := \mathbb{E}_{X\sim p^\star}\Big[D_{\text{KL}}\big(p_i^\star(\cdot \mid X_W) \,\|\, p_{\theta,i}(\cdot \mid X_W)\big)\Big].$$

By the KL chain rule in reverse generation order, we have

$$\begin{aligned} D_{\text{KL}}(p^\star \| p_\theta^{\mathcal{S}}) &= \sum_{t=1}^T \mathbb{E}_{X\sim p^\star}\Big[D_{\text{KL}}\Big(p^\star(X_{S^t} \in \cdot \mid X_{W^t}) \,\Big\|\, \bigotimes_{i\in S^t} p_{\theta,i}(\cdot \mid X_{W^t})\Big)\Big] \\ &= \sum_{t=1}^T \text{TC}_{p^\star}(X_{S^t} \mid X_{W^t}) + \sum_{t=1}^T \sum_{i\in S^t} \Delta_{\theta,i}(W^t). \end{aligned}$$

Here the second equality follows by adding and subtracting the log of the product of the true conditional marginals. Averaging the first term over $\mathcal{S}$ gives $\varepsilon_{\text{parallel}}(p^\star; \boldsymbol{s})$ by (30).

For the second term, let $M^t := [L] \setminus W^t$ and $m_t := \sum_{u=1}^t s_u$. Then M^t is uniform among subsets of $[L]$ of size m_t, and, conditioned on M^t, the set S^t is a uniform s_t-subset of M^t. Hence,

$$\begin{aligned} \mathbb{E}_{\mathcal{S}}\Big[\sum_{t=1}^T \sum_{i\in S^t} \Delta_{\theta,i}(W^t)\Big] &= \sum_{t=1}^T \frac{s_t}{m_t}\, \mathbb{E}_{M^t}\Big[\sum_{i\in M^t} \Delta_{\theta,i}([L]\setminus M^t)\Big] \\ &= L\mathcal{L}(\theta). \end{aligned}$$

The last equality uses the definition of the centered loss $\mathcal{L}$, with training time distribution $\mathbb{P}[\tau = t] = s_t/L$. We therefore obtain the decomposition

$$\mathbb{E}_{\mathcal{S}}\big[D_{\text{KL}}(p^\star \| p_\theta^{\mathcal{S}})\big] = \varepsilon_{\text{parallel}}(p^\star; \boldsymbol{s}) + L\mathcal{L}(\theta). \tag{39}$$

Finally, taking expectation over the training sample and applying Theorem 1 gives

$$\begin{aligned} \mathbb{E}_{\{X^{(j)}\},\mathcal{S}}\big[D_{\text{KL}}(p^\star \| p_\theta^{\mathcal{S}})\big] &= \varepsilon_{\text{parallel}}(p^\star; \boldsymbol{s}) + L\mathbb{E}_{\{X^{(j)}\}}\big[\mathcal{L}(\widehat{\theta})\big] \\ &\lesssim \varepsilon_{\text{parallel}}(p^\star; \boldsymbol{s}) + s_{\max}\mathsf{Proj}_{J,R}(p^\star) + \frac{LK_{\text{par}}}{n}\,\mathsf{poly}\log(n, L, V, J, R). \end{aligned}$$

This proves the theorem.

B Proofs of Propositions and lemmas

B.1 Lemma 4 and its proof

Lemma 4 (Per-token denoising loss is average conditional KL)**.** *Recall the definition of τ, M^τ and W^τ in Section 2.1. Let random index I follows uniform distribution on M^τ conditioned on (τ, M^τ). Loss function $\mathcal{L}(\theta)$ satisfies*

$$\begin{aligned}\mathcal{L}(\theta) &:= \overline{R}(p_\theta) - \overline{R}(p^\star) \\ &= \mathbb{E}_{\tau, M^\tau, I,\, X\sim p^\star}\Big[D_{\mathrm{KL}}\big(p_I^\star(\cdot \mid X_{W^\tau}) \,\|\, p_{\theta,I}(\cdot \mid X_{W^\tau})\big)\Big]. \end{aligned} \tag{40}$$

Here $\overline{R}(p_\theta) := \mathbb{E}_{X\sim p^\star}[\ell_{p_\theta}(X)]$ is the risk of the conditional predictor p_θ.

Proof of Lemma 4. Subtract the cross-entropy of the true conditional predictor from that of the network and introduce a conditionally uniform $I \sim \mathrm{Unif}(M^\tau)$. Then

$$\begin{aligned}\overline{\mathcal{R}}(p_\theta) - \overline{\mathcal{R}}(p^\star) &= \mathbb{E}_{\tau, M^\tau, X\sim p^\star}\Big[\frac{1}{|M^\tau|}\sum_{i\in M^\tau} \log \frac{p_i^\star(X_i \mid X_{W^\tau})}{p_{\theta,i}(X_i \mid X_{W^\tau})}\Big] \\ &= \mathbb{E}_{\tau, M^\tau, I, X\sim p^\star}\Big[\log \frac{p_I^\star(X_I \mid X_{W^\tau})}{p_{\theta,I}(X_I \mid X_{W^\tau})}\Big].\end{aligned}$$

Conditional on (X_{W^τ}, I), the token distribution X_I follows the law $p_I^\star(\cdot \mid X_{W^\tau})$. The inner conditional expectation is precisely the KL divergence in (40).

B.2 Details of Proposition 1 and its proof

Denote the infinity-norm distance for functions on $\overline{\mathbb{V}}^L$ as,

$$d_\infty(F, F') := \sup_{c\in\overline{\mathbb{V}}^L} \max_{i\in[L], v\in\mathbb{V}} |F_{i,v}(c) - F'_{i,v}(c)|. \tag{41}$$

Write $\mathcal{N}(\eta, \mathcal{F}, d_\infty)$ for the minimum number of radius-η balls needed to cover $\mathcal{F}$.

Assumption 1 (Estimation class)**.** The logit class $\mathcal{F}(\Theta)$ satisfies:

(i) Boundedness: $-G \le F_{\theta,i,k}(x) \le 0$ for every allowed input, row, token, and parameter. Hence after softmax operator the output predictor $p_{\theta,i}(k \mid c) \ge e^{-G}/V$.

(ii) Capacity: for some deterministic $C(\Theta) \ge 1$ and every $0 < \eta \le 1$,

$$\log \mathcal{N}(\eta, \mathcal{F}(\Theta), d_\infty) \le C(\Theta)\log(e/\eta).$$

Proof of Proposition 1. In this section, we prove the localized oracle inequality. Throughout the proof, the masking schedule $\boldsymbol{s}$ is fixed, and C denotes a universal constant that may change from line to line.

For a function g of a token sequence, write

$$Pg := \mathbb{E}_{X\sim p^\star}[g(X)], \qquad P_n g := \frac{1}{n}\sum_{j=1}^n g(X^{(j)}).$$

Recall that ℓ_{p_θ} already averages over the masking randomness. Hence, P_n averages over the n independent training sequences. For $\zeta > 0$, choose $\theta^\circ \in \Theta$ such that

$$a := \mathcal{L}(\theta^\circ) \le \varepsilon_{\mathrm{approx}}(\Theta) + \zeta, \qquad h_\theta := \ell_{p_\theta} - \ell_{p_{\theta^\circ}}.$$

By the definition of the centered loss and the ERM property,

$$Ph_\theta = \mathcal{L}(\theta) - a, \qquad P_n h_{\widehat{\theta}} \le 0.$$

We first bound the second moment of h_θ and then control its empirical deviation.

Second-moment bound. Let $M_G = G + \log V$ and $\rho = e^{-M_G}$. The bounded logits imply $p_{\theta,i}(v \mid x) \geq \rho$ for every masked position and token. We claim that, for any probability vectors p, q on $[V]$ with $q_v \geq \rho$,

$$\sum_{v:p_v>0} p_v \log^2 \frac{p_v}{q_v} \leq (M_G+2) D_{\mathrm{KL}}(p\|q) \leq 3M_G D_{\mathrm{KL}}(p\|q). \tag{42}$$

Here the last inequality uses $M_G \geq \log 3 > 1$. To prove the first inequality, notice that

$$t\log^2 t \leq (M_G+2)(t\log t - t + 1), \qquad \forall 0 \leq t \leq e^{M_G}.$$

Indeed, for $t = e^{-x} \leq 1$, the inequality $e^x \geq 1 + x + x^2/2$ gives

$$e^{-x}x^2 \leq 2 \cdot \big(1 - (x+1)e^{-x}\big).$$

For $t = e^x \geq 1$, with $0 \leq x \leq M_G$, we have

$$(x+2)(x-1+e^{-x}) - x^2 = x - 2 + (x+2)e^{-x} \geq 0,$$

since the last expression is zero at $x = 0$ and has derivative $1 - (x+1)e^{-x} \geq 0$. The case $t = 0$ follows by continuity. Taking $t = p_v/q_v$, multiplying by q_v, and summing over v proves (42).

Since $0 \leq \ell_{p_\theta} \leq M_G$, we have $\|h_\theta\|_\infty \leq M_G$. Let τ and M_τ follow the training masking distribution, set $W^\tau = [L] \setminus M_\tau$, and let I be uniform in M_τ. Jensen's inequality gives

$$\begin{aligned} Ph_\theta^2 &\leq \mathbb{E}\left[\log^2 \frac{p_{\theta^\circ,I}(X_I \mid X_{W^\tau})}{p_{\theta,I}(X_I \mid X_{W^\tau})}\right] \\ &\leq 2\mathbb{E}\left[\log^2 \frac{p^\star_I(X_I \mid X_{W^\tau})}{p_{\theta,I}(X_I \mid X_{W^\tau})}\right] + 2\mathbb{E}\left[\log^2 \frac{p^\star_I(X_I \mid X_{W^\tau})}{p_{\theta^\circ,I}(X_I \mid X_{W^\tau})}\right]. \end{aligned}$$

Here the expectation includes $X \sim p^\star$ and (τ, M_τ, I). Conditioned on the context and I, the token X_I has law $p^\star_I(\cdot \mid X_{W^\tau})$. Applying (42) and the conditional KL expression for $\mathcal{L}$ therefore yields

$$Ph_\theta^2 \leq 6M_G\big(\mathcal{L}(\theta) + a\big). \tag{43}$$

Control of the empirical deviation. Let F_θ denote the clipped logit function on masked sequences, including the input embedding. By the Lipschitz property of log-softmax,

$$\|h_\theta - h_{\theta'}\|_\infty = \|\ell_{p_\theta} - \ell_{p_{\theta'}}\|_\infty \leq 2d_\infty(F_\theta, F_{\theta'}).$$

For $r > 0$, define

$$\mathcal{H}(r) := \{h_\theta : \theta \in \Theta,\ \mathcal{L}(\theta) \leq r\}.$$

The covering bound in Assumption 1 then implies

$$\log \mathcal{N}(\eta, \mathcal{H}(r), \|\cdot\|_\infty) \leq C(\Theta)\log\frac{8e}{\eta}, \qquad 0 < \eta \leq 1,$$

whenever $\mathcal{H}(r)$ is nonempty. To take the covering centers in $\mathcal{H}(r)$, choose one member from each ball intersecting this class and double the radius.

Let $a_n := C(\Theta)\log(cn)$ for a sufficiently large universal constant $c \geq 2$. Applying Lemma 8 with $b = M_G$ and $\sigma^2 = 6M_G(r+a)$ gives, for fixed $r, t > 0$,

$$\sup_{h\in\mathcal{H}(r)} (P - P_n)h \leq C\left[\sqrt{\frac{M_G(r+a)(a_n+t)}{n}} + \frac{M_G(a_n+t)}{n}\right] \tag{44}$$

with probability at least $1 - e^{-t}$.

We apply this bound at a sequence of increasing risk levels. Fix $u > 0$ and set

$$r_0 := 4a + AM_G\frac{a_n+u}{n}, \qquad r_j := 2^j r_0, \quad j \geq 1,$$

where A is a sufficiently large universal constant. Taking $r = r_j$ and $t = u + j\log 2$ in (44), and using $\sum_{j\geq 1} e^{-u-j\log 2} = e^{-u}$, all these bounds hold simultaneously with probability at least $1 - e^{-u}$. Moreover,

$$r_j + a \leq 3r_{j-1}, \quad a_n + u + j\log 2 \leq C2^{j-1}(a_n+u), \quad \frac{M_G(a_n+u+j\log 2)}{n} \leq \frac{C}{A}r_{j-1}.$$

Consequently, on the same event,

$$\sup_{h\in\mathcal{H}(r_j)} (P-P_n)h \le C(A^{-1/2}+A^{-1})r_{j-1} \le \frac{1}{2}\, r_{j-1},$$

after choosing A large enough.

Suppose that $\mathcal{L}(\widehat{\theta}) > r_0$, and choose $j \ge 1$ such that $r_{j-1} < \mathcal{L}(\widehat{\theta}) \le r_j$. Since $a \le r_{j-1}/4$, the ERM property gives

$$\frac{3}{4} r_{j-1} < \mathcal{L}(\widehat{\theta}) - a = P h_{\widehat{\theta}} \le (P-P_n) h_{\widehat{\theta}} \le \frac{1}{2} r_{j-1},$$

a contradiction. Thus, with probability at least $1-e^{-u}$,

$$\mathcal{L}(\widehat{\theta}) \le 4a + C M_G \frac{C(\Theta)\log(en)+u}{n}.$$

Using $a \le \varepsilon_{\mathrm{approx}}(\Theta)+\zeta$ and letting $\zeta \downarrow 0$, we obtain

$$\mathbb{P}\left[\mathcal{L}(\widehat{\theta}) > 4\varepsilon_{\mathrm{approx}}(\Theta) + C M_G \frac{C(\Theta)\log(en)+u}{n}\right] \le e^{-u}. \tag{45}$$

Finally, integrating (45) over $u>0$ yields

$$\begin{aligned}\mathbb{E}\big[\mathcal{L}(\widehat{\theta})\big] &\le 4\varepsilon_{\mathrm{approx}}(\Theta) + \frac{C M_G}{n}\left[C(\Theta)\log(en) + \int_0^\infty e^{-u}\,\mathrm{d}u\right] \\ &= 4\varepsilon_{\mathrm{approx}}(\Theta) + C M_G \cdot \frac{C(\Theta)\log(en)+1}{n}.\end{aligned}$$

This proves Proposition 1.

B.3 Lemma 5 and its proof

We bound the capacity of the clipped logit class associated with the predictor in (9). For $\theta = (E, P, \theta_{\mathrm{tf}}, W_{\mathrm{out}}) \in \Theta$, denote

$$F_\theta(x) := \mathrm{Clip}_{[-G,0]}\big(\mathsf{TF}_{\theta_{\mathrm{tf}}}(\mathsf{Embed}_{E,P}(x))W_{\mathrm{out}}\big), \qquad \mathcal{F}(\Theta) := \{F_\theta : \theta \in \Theta\}.$$

Thus, $p_{\theta,i}(\cdot \mid x) = \mathrm{softmax}(F_{\theta,i,\cdot}(x))$. Recall the definition of $d_\infty(F,F')$ in (41).

Lemma 5 (Capacity of the sparse Transformer class)**.** *Let $\Theta = \Theta(D,N,M,D_{\mathrm{ff}},B,S;R_{\mathrm{embed}},G)$ be the parameter class in Section 2.2, with $D,N,M,D_{\mathrm{ff}} \ge 1$ and $B,R_{\mathrm{embed}},G>0$. Define*

$$P_{\mathrm{tr}} := (V+1)D + LD + N(3MD^2+2DD_{\mathrm{ff}}) + DV = O(DV+LD+NMD^2+NDD_{\mathrm{ff}}),$$

and suppose the sparsity $1 \le S \le P_{\mathrm{tr}}$ is an integer. Denote $A_{\mathrm{TF}} \ge e$ satisfying

$$\log A_{\mathrm{TF}} \le CN^2 \log\big(eNM(B\vee 1)(R_{\mathrm{embed}}\vee 1)\big)$$

such that, for every $0<\eta\le 1$,

$$\log \mathcal{N}(\eta, \mathcal{F}(\Theta), d_\infty) \le S\log\frac{eP_{\mathrm{tr}}}{S} + S\log\frac{A_{\mathrm{TF}}}{\eta}. \tag{46}$$

Consequently, Assumption 1 holds with

$$C(\Theta) = S\left[\log\frac{eP_{\mathrm{tr}}}{S} + \log A_{\mathrm{TF}}\right].$$

The same covering bound holds for the probability-valued predictor class in the uniform distance. If $S > P_{\mathrm{tr}}$, replace S by P_{tr} in these bounds.

Proof of Lemma 5. We first control the change in the logits under a perturbation of the parameters, and then cover the sparse parameter space. Write $\|Z\|_{2,\infty} := \max_i \|Z_{i,\cdot}\|_2$, and let $d_\Theta(\theta, \theta')$ be the largest Frobenius distance between corresponding parameter matrices, including E, P, and W_{out}. Set

$$\overline{B} := B \vee 1, \qquad \overline{R} := R_{\mathrm{embed}} \vee 1, \qquad a := eNM\overline{B}\,\overline{R}.$$

For $\|Z\|_{2,\infty} \le R$, the bounded attention activation gives

$$\|\mathrm{head}_\theta(Z)_{i,\cdot}\|_2 \le \frac{1}{L}\sum_{j=1}^{L} \big|\sigma_{\mathrm{attn}}(\langle Z_{i,\cdot}Q, Z_{j,\cdot}K\rangle)\big| \|Z_{j,\cdot}V\|_2 \le BR.$$

Hence the attention layer maps the row radius R to at most $(1 + MB)R$. The feed-forward layer multiplies this radius by at most $1+B^2$, since ReLU is one-Lipschitz and vanishes at zero. Denoting the hidden states by $H_\theta^{(\ell)}(x)$, with $H_\theta^{(0)}(x) = \mathsf{Embed}_{E,P}(x)$, we obtain

$$\|H_\theta^{(\ell)}(x)\|_{2,\infty} \le R_{\mathrm{embed}}\big[(1 + MB)(1 + B^2)\big]^\ell \le a^{cN}, \qquad 0 \le \ell \le N,$$

uniformly over x and $\theta \in \Theta$.

Next, let $\Delta := d_\Theta(\theta, \theta')$. The trainable embeddings satisfy

$$\|H_\theta^{(0)}(x) - H_{\theta'}^{(0)}(x)\|_{2,\infty} \le \|E - E'\|_{\mathrm{F}} + \|P - P'\|_{\mathrm{F}} \le 2\Delta.$$

For one attention head, the sigmoid activation is $1/4$-Lipschitz. For inputs Z, Z' with row norms at most R, we have

$$\begin{aligned}
\|\mathrm{head}_\theta(Z) - \mathrm{head}_\theta(Z')\|_{2,\infty} &\le \big(B + \tfrac{1}{2}B^3R^2\big)\|Z - Z'\|_{2,\infty},\\
\|\mathrm{head}_\theta(Z) - \mathrm{head}_{\theta'}(Z)\|_{2,\infty} &\le \big(R + \tfrac{1}{2}B^2R^3\big)\Delta.
\end{aligned}$$

For the feed-forward layer, the corresponding bounds are

$$\begin{aligned}
\|\mathsf{FFN}_\theta(Z) - \mathsf{FFN}_\theta(Z')\|_{2,\infty} &\le (1 + B^2)\|Z - Z'\|_{2,\infty},\\
\|\mathsf{FFN}_\theta(Z) - \mathsf{FFN}_{\theta'}(Z)\|_{2,\infty} &\le 2BR\Delta.
\end{aligned}$$

Combining these bounds with the hidden-state bound gives

$$\|H_\theta^{(\ell)}(x) - H_{\theta'}^{(\ell)}(x)\|_{2,\infty} \le a^{cN}\big(\|H_\theta^{(\ell-1)}(x) - H_{\theta'}^{(\ell-1)}(x)\|_{2,\infty} + \Delta\big).$$

Starting from the embedding perturbation bound and iterating over N blocks, we obtain

$$\|H_\theta^{(N)}(x) - H_{\theta'}^{(N)}(x)\|_{2,\infty} \le a^{cN^2}\Delta,$$

after increasing the universal constant c.

For the output embedding, let $\|\cdot\|_{\max}$ denote the largest absolute matrix entry. Then

$$\begin{aligned}
&\|H_\theta^{(N)}(x)W_{\mathrm{out}} - H_{\theta'}^{(N)}(x)W'_{\mathrm{out}}\|_{\max}\\
&\quad\le B\|H_\theta^{(N)}(x) - H_{\theta'}^{(N)}(x)\|_{2,\infty} + \|H_{\theta'}^{(N)}(x)\|_{2,\infty}\|W_{\mathrm{out}} - W'_{\mathrm{out}}\|_{\mathrm{F}}\\
&\quad\le a^{cN^2}\Delta.
\end{aligned}$$

Since clipping is one-Lipschitz, taking the supremum over masked sequences yields

$$d_\infty(F_\theta, F_{\theta'}) \le \Lambda d_\Theta(\theta, \theta'), \qquad \Lambda := a^{cN^2}.$$

It remains to cover the parameters. The input and positional embeddings have $(V + 1)D + LD$ entries, each Transformer block has $3MD^2 + 2DD_{\mathrm{ff}}$ entries, and the output embedding has DV entries. This gives the stated value of P_{tr}.

For a fixed support $\mathcal{I} \subseteq [P_{\mathrm{tr}}]$ of size S, denote

$$\Theta_{\mathcal{I}} := \{\theta \in \Theta : \mathsf{supp}(\theta) \subseteq \mathcal{I}\}.$$

Ignoring the embedded-input constraint temporarily, the parameters on this support lie in a product of Frobenius balls of radius B, with total dimension S. Covering each ball at radius $\zeta/2$ gives a

product cover of size at most $(1 + 4B/\zeta)^S$ in the maximum Frobenius distance. For each ball intersecting $\Theta_{\mathcal{I}}$, choose one parameter from the intersection as its new center. The radius is then at most ζ, and every center satisfies all constraints, including the embedded-input bound. Consequently,

$$\log \mathcal{N}(\zeta, \Theta_{\mathcal{I}}, d_\Theta) \le S \log\left(1 + \frac{4B}{\zeta}\right).$$

Every parameter with at most S nonzero entries belongs to one of these support classes. Taking the union over supports gives

$$\begin{aligned}\log \mathcal{N}(\zeta, \Theta, d_\Theta) &\le \log \binom{P_{\mathrm{tr}}}{S} + S \log\left(1 + \frac{4B}{\zeta}\right) \\ &\le S \log \frac{eP_{\mathrm{tr}}}{S} + S \log\left(1 + \frac{4B}{\zeta}\right).\end{aligned}$$

Take $\zeta = \eta/\Lambda$ and $A_{\mathrm{TF}} := e(1 + 4B\Lambda)$. Transferring the parameter cover to the logits proves (46), and

$$\log A_{\mathrm{TF}} \le CN^2 \log\big(eNM(B \vee 1)(R_{\mathrm{embed}} \vee 1)\big).$$

Since $A_{\mathrm{TF}} \ge e$ and $0 < \eta \le 1$, the covering bound is at most $C(\Theta)\log(e/\eta)$. The logits already lie in $[-G, 0]$, so both parts of Assumption 1 hold.

Finally, softmax is one-Lipschitz in the maximum norm. Applying it rowwise therefore transfers the same cover to the predictor class. This completes the proof.

B.4 Regularization and its property

In this section, we regularize distributions while preserving its tree-mixture structure. This gives positive conditional probabilities and bounded conditional log probabilities for the following transformer approximation.

Definition 4 (δ-regularization)**.** For $0 < \delta < 1$, let $Q_\delta(a, b) := (1 - \delta)\mathbf{1}\{a = b\} + \delta/V$ as the transition kernel. For any distribution q on $\mathbb{V}^L$, define its δ-regularization by

$$q^\delta(y) := (P_\delta q)(y) := \sum_{x \in \mathbb{V}^L} q(x) \prod_{i=1}^{L} Q_\delta(x_i, y_i), \qquad y \in \mathbb{V}^L. \tag{47}$$

This transition independently replaces each token by a uniform token with probability δ.

Lemma 6 (Properties of δ-regularization)**.** *Let $0 < \delta < 1$, $\rho_\delta := \delta/V$, and $G_\delta := \log(V/\delta)$. The following properties hold.*

(i) *If $q = \sum_{r=1}^{R} \pi_r q_r \in \mathcal{M}_{J,R}$, then*

$$q^\delta = \sum_{r=1}^{R} \pi_r q_r^\delta \in \mathcal{M}_{J,R}, \quad B_{i,r}^\delta(v, z) = (1 - \delta)B_{i,r}(v, z) + \frac{\delta}{V}. \tag{48}$$

Here $q_r^\delta = P_\delta q_r$ is obtained by replacing the emission tables $B_{i,r}$ with $B_{i,r}^\delta$ as in (11)*. The root distributions, transitions, component trees, and mixture weights are unchanged.*

(ii) *For any distribution q on $\mathbb{V}^L$, any $y \in \mathbb{V}^L$, and any masked sequence $c \in \overline{\mathbb{V}}^L$, it holds that:*

$$q^\delta(y) \ge \rho_\delta^L, \quad q^\delta(y) \ge (1 - \delta)^L q(y), \quad q_i^\delta(v \mid c) \ge \rho_\delta, \quad \forall i \notin W(c), \text{ and } v \in \mathbb{V}. \tag{49}$$

Here $W(c) := \{j \in [L] : c_j \ne \mathtt{M}\}$ and $q_i^\delta(v \mid c) := \mathbb{P}_{X \sim q^\delta}[X_i = v \mid X_{W(c)} = c_{W(c)}]$.

(iii) *For any distributions p, q on $\mathbb{V}^L$,*

$$D_{\mathrm{KL}}(p \| q^\delta) \le D_{\mathrm{KL}}(p \| q) + L \log \frac{1}{1 - \delta}. \tag{50}$$

Proof of Lemma 6. For part (i), fix a representation $q = \sum_{r=1}^{R} \pi_r q_r$ as in Section 3. Notice that

$$\sum_{a \in \mathbb{V}} B_{i,r}(a, z) Q_\delta(a, v) = (1 - \delta) B_{i,r}(v, z) + \frac{\delta}{V} = B^\delta_{i,r}(v, z).$$

Substituting the hidden-tree factorization into (47) gives

$$\begin{aligned}(P_\delta q_r)(y) &= \sum_{z \in [J]^L} \nu_r(z_{o_r}) \prod_{i \neq o_r} A_{i,r}(z_i, z_{\mathrm{pa}_r(i)}) \prod_{i=1}^{L} \Big[\sum_{a \in \mathbb{V}} B_{i,r}(a, z_i) Q_\delta(a, y_i) \Big] \\ &= \sum_{z \in [J]^L} \nu_r(z_{o_r}) \prod_{i \neq o_r} A_{i,r}(z_i, z_{\mathrm{pa}_r(i)}) \prod_{i=1}^{L} B^\delta_{i,r}(y_i, z_i) = q^\delta_r(y). \end{aligned} \tag{51}$$

Thus, regularization changes only the emission tables. By linearity, $q^\delta = \sum_{r=1}^{R} \pi_r q^\delta_r \in \mathcal{M}_{J,R}$. Since (47) is defined using the observed law q, the resulting law does not depend on the chosen hidden-tree representation.

For part (ii), the bounds $Q_\delta(a, b) \geq \rho_\delta$ and $Q_\delta(a, a) \geq 1 - \delta$ imply

$$q^\delta(y) \geq \sum_{x \in \mathbb{V}^L} q(x) \rho_\delta^L = \rho_\delta^L, \tag{52}$$

$$q^\delta(y) \geq q(y) \prod_{i=1}^{L} Q_\delta(y_i, y_i) \geq (1 - \delta)^L q(y). \tag{53}$$

Fix a masked sequence c and write $W = W(c)$. Summing over the coordinates outside W gives

$$\mathbb{P}_{X \sim q^\delta}[X_W = c_W] = \sum_{x \in \mathbb{V}^L} q(x) \prod_{j \in W} Q_\delta(x_j, c_j) \geq \rho_\delta^{|W|} > 0.$$

Hence, for $i \notin W$ and $v \in \mathbb{V}$,

$$q^\delta_i(v \mid c) = \frac{\sum_{x \in \mathbb{V}^L} q(x) Q_\delta(x_i, v) \prod_{j \in W} Q_\delta(x_j, c_j)}{\sum_{x \in \mathbb{V}^L} q(x) \prod_{j \in W} Q_\delta(x_j, c_j)} \geq \rho_\delta.$$

For part (iii), the result is immediate if $D_{\mathrm{KL}}(p \| q) = +\infty$. Otherwise, applying (53) yields

$$\begin{aligned} D_{\mathrm{KL}}(p \| q^\delta) - D_{\mathrm{KL}}(p \| q) &= \sum_{y : p(y) > 0} p(y) \log \frac{q(y)}{q^\delta(y)} \\ &\leq L \log \frac{1}{1 - \delta}. \end{aligned}$$

This proves the lemma. In particular, regularization also preserves any restriction on the component tree topologies, including a prescribed height bound.

B.5 Decomposition of the approximation error based on regularization

In this section, we bound the approximation error by the joint KL error of a comparison distribution, the regularization error, and the network approximation error.

Fix $0 < \delta < 1$ and a distribution q on $\mathbb{V}^L$. For a masked sequence $c \in \overline{\mathbb{V}}^L$, define the target logits of its δ-regularized distribution as in Definition 4 by

$$F^{\delta,q}_{i,v}(c) := \begin{cases} \log q^\delta_i(v \mid c), & i \notin W(c), \\ 0, & i \in W(c), \end{cases} \quad \forall i \in [L], \forall v \in \mathbb{V}. \tag{54}$$

By Lemma 6, these logits lie in $[-G_\delta, 0]$, and

$$\mathrm{softmax}(F^{\delta,q}_{i,\cdot}(c)) = q^\delta_i(\cdot \mid c), \qquad i \notin W(c).$$

The values at unmasked positions are set to zero only to define a full $L \times V$ matrix; they do not enter the loss or the approximation distance below.

Recall the clipped logit class $\mathcal{F}(\Theta)$ in Appendix B.3. Define its uniform approximation error on masked positions by

$$d_{\infty,\mathrm{mask}}(F^{\delta,q}, \mathcal{F}(\Theta)) := \inf_{\theta\in\Theta} \sup_{\substack{c\in\overline{\mathbb{V}}^L \\ W(c)\neq[L]}} \max_{\substack{i\notin W(c) \\ v\in\mathbb{V}}} |F_{\theta,i,v}(c) - F^{\delta,q}_{i,v}(c)|. \tag{55}$$

Lemma 7 (Decomposition of the approximation error)**.** *Fix $0 < \delta < 1$ and a fixed-cardinality masking schedule $\boldsymbol{s} = (s_1, \ldots, s_T)$, with $s_t \in \mathbb{N}_+$ and $\sum_{t=1}^T s_t = L$. For any distribution q on $\mathbb{V}^L$,*

$$\varepsilon_{\mathrm{approx}}(\Theta) \le \frac{s_{\max}}{L} D_{\mathrm{KL}}(p^\star \| q^\delta) + 2 d_{\infty,\mathrm{mask}}(F^{\delta,q}, \mathcal{F}(\Theta)), \tag{56}$$

where $s_{\max} := \max_t s_t$ and $\varepsilon_{\mathrm{approx}}(\Theta) := \inf_{\theta\in\Theta} \mathcal{L}(\theta)$. In particular, if function class $\mathcal{F}(\Theta)$ satisfies

$$\sup_{q\in\mathcal{M}_{J,R}} d_{\infty,\mathrm{mask}}(F^{\delta,q}, \mathcal{F}(\Theta)) \le \epsilon,$$

then

$$\varepsilon_{\mathrm{approx}}(\Theta) \le \frac{s_{\max}}{L} \mathsf{Proj}_{J,R}(p^\star) + s_{\max} \log \frac{1}{1-\delta} + 2\epsilon. \tag{57}$$

The same conclusion holds for any nonempty subclass $\mathcal{Q} \subseteq \mathcal{M}_{J,R}$, with the uniform approximation condition imposed on $\mathcal{Q}$ and $\mathsf{Proj}_{J,R}$ replaced by $\mathsf{Proj}_{\mathcal{Q}}$.

Proof of Lemma 7. Fix q and write $r = q^\delta$. Let r_{cond} denote its conditional predictor on masked positions. We use the same training experiment as in Section 2: $\mathbb{P}[\tau = t] = s_t/L$, M^τ is uniform among subsets of size $m_\tau := \sum_{u=1}^\tau s_u$, and I is uniform in M^τ. Write $W^\tau = [L] \setminus M^\tau$, and let C^τ be the masked sequence with entries X_j for $j \in W^\tau$ and $\mathtt{M}$ otherwise, where $X \sim p^\star$. Define

$$\mathcal{L}(r_{\mathrm{cond}}) := \mathbb{E}\big[D_{\mathrm{KL}}\big(p^\star_I(\cdot \mid C^\tau) \,\|\, r_I(\cdot \mid C^\tau)\big)\big].$$

Here the mask draws are independent of X. By Lemma 4, adding and subtracting the log conditional probability under r gives

$$\begin{aligned}\mathcal{L}(\theta) &= \mathbb{E}\left[\log \frac{p^\star_I(X_I \mid C^\tau)}{p_{\theta,I}(X_I \mid C^\tau)}\right] \\ &= \mathcal{L}(r_{\mathrm{cond}}) + \mathbb{E}\left[\log \frac{r_I(X_I \mid C^\tau)}{p_{\theta,I}(X_I \mid C^\tau)}\right].\end{aligned} \tag{58}$$

For any $f, g \in \mathbb{R}^V$, the log-sum-exp function is one-Lipschitz in the maximum norm. Hence,

$$\begin{aligned}|\log \mathrm{softmax}_v(f) - \log \mathrm{softmax}_v(g)| &\le |f_v - g_v| + \left|\log \sum_{a=1}^V e^{f_a} - \log \sum_{a=1}^V e^{g_a}\right| \\ &\le 2\|f - g\|_\infty.\end{aligned} \tag{59}$$

Since $r_i(\cdot \mid c) = \mathrm{softmax}(F^{\delta,q}_{i,\cdot}(c))$, applying this bound in (58) and taking the infimum over $\theta \in \Theta$ yields

$$\varepsilon_{\mathrm{approx}}(\Theta) \le \mathcal{L}(r_{\mathrm{cond}}) + 2 d_{\infty,\mathrm{mask}}(F^{\delta,q}, \mathcal{F}(\Theta)).$$

It remains to control $\mathcal{L}(r_{\mathrm{cond}})$ by $D_{\mathrm{KL}}(p^\star \| r)$. We use a subset-averaging argument, analogous to the entropy calculation in (Chen et al., 2025, proof of Theorem 3.3). For $W \subseteq [L]$, denote by $p^\star_W$ and r_W the corresponding marginals. Let

$$\overline{d}_j := \binom{L}{j}^{-1} \sum_{\substack{W\subseteq[L] \\ |W|=j}} D_{\mathrm{KL}}(p^\star_W \| r_W), \qquad 0 \le j \le L,$$

and write $\Delta_j := \overline{d}_{j+1} - \overline{d}_j$ for $0 \le j < L$. By the KL chain rule, for $i \notin W$,

$$D_{\mathrm{KL}}(p^\star_{W\cup\{i\}} \| r_{W\cup\{i\}}) - D_{\mathrm{KL}}(p^\star_W \| r_W) = \mathbb{E}_{X_W\sim p^\star_W}\big[D_{\mathrm{KL}}\big(p^\star_i(\cdot \mid X_W) \,\|\, r_i(\cdot \mid X_W)\big)\big].$$

If W is a uniform j-subset and I is uniform in $[L] \setminus W$, then $W \cup \{I\}$ is a uniform $(j+1)$-subset. Averaging the preceding identity gives

$$\mathbb{E}_{W,I,X_W}\left[D_{\mathrm{KL}}\big(p_I^\star(\cdot \mid X_W) \,\|\, r_I(\cdot \mid X_W)\big)\right] = \Delta_j \geq 0. \tag{60}$$

All these divergences are finite because r has full support. Since $\overline{d}_0 = 0$ and $\overline{d}_L = D_{\mathrm{KL}}(p^\star \| r)$,

$$\sum_{j=0}^{L-1} \Delta_j = D_{\mathrm{KL}}(p^\star \| r). \tag{61}$$

At training time t, the context set W^t is uniform of size $L - m_t$, where $m_t = \sum_{u=1}^{t} s_u$. Therefore,

$$\begin{aligned}\mathcal{L}(r_{\mathrm{cond}}) &= \frac{1}{L}\sum_{t=1}^{T} s_t \Delta_{L-m_t} \\ &\leq \frac{s_{\max}}{L}\sum_{j=0}^{L-1}\Delta_j = \frac{s_{\max}}{L} D_{\mathrm{KL}}(p^\star \| r).\end{aligned} \tag{62}$$

Here we use $\Delta_j \geq 0$ and the fact that the indices $L - m_t$ are distinct, since $s_t \geq 1$. Combining this with the network approximation bound proves (56).

Under the uniform approximation condition, Lemma 6(iii) further gives, for every $q \in \mathcal{M}_{J,R}$,

$$\begin{aligned}\varepsilon_{\mathrm{approx}}(\Theta) &\leq \frac{s_{\max}}{L} D_{\mathrm{KL}}(p^\star \| q^\delta) + 2\epsilon \\ &\leq \frac{s_{\max}}{L} D_{\mathrm{KL}}(p^\star \| q) + s_{\max} \log \frac{1}{1-\delta} + 2\epsilon.\end{aligned}$$

Taking the infimum over $q \in \mathcal{M}_{J,R}$ proves (57). The same argument applies when this infimum is restricted to $\mathcal{Q}$.

The bound (57) is the approximation bound used in Proposition 2 of the main text. Thus, the remaining construction only needs to approximate $F^{\delta,q}$ uniformly over $q \in \mathcal{M}_{J,R}$ with one common architectural budget. For the shallow-tree result, the same argument uses $\mathcal{Q} = \mathcal{M}_{J,R}^h$ and the corresponding projection error.

B.6 A LOCALIZED EMPIRICAL-PROCESS LEMMA

Lemma 8 (Localized bounds from a uniform cover)**.** *Let $\mathcal{G}$ be a nonempty function class with $\|g\|_\infty \leq b$ and $Pg^2 \leq \sigma^2$ for every $g \in \mathcal{G}$, where $b \geq 1$. Suppose $v \geq 1$, $A \geq e$, and its covering number control*

$$\log \mathcal{N}(\eta, \mathcal{G}, \|\cdot\|_\infty) \leq v \log(A/\eta), \quad \forall\, 0 < \eta \leq 1.$$

For $n \geq 2$, take $d_n := v \log(cAn)$, for a universal $c \geq 2$. Then for $t > 0$, with probability at least $1 - e^{-t}$,

$$\sup_{g \in \mathcal{G}} (P - P_n) g \leq C\Big(\sigma\sqrt{\frac{d_n + t}{n}} + \frac{b(d_n + t)}{n}\Big).$$

Proof of Lemma 8. Take an internal $1/n$-net $\{g_1, \ldots, g_J\} \subseteq \mathcal{G}$. If covering centers are initially outside the class, start with a $1/(2n)$-cover and choose one element of each intersecting ball. Thus $\log(2J) \leq v \log(4An) \leq d_n$ after increasing c. All centers retain the variance and envelope bounds.

For the deviation assertion, the independent centered variables $Pg_j - g_j(X^{(i)})$ have magnitude at most $2b$ and variance at most σ^2. Bernstein's inequality and a union bound then give

$$\max_{1 \leq j \leq J} (P - P_n) g_j \leq \sqrt{\frac{2\sigma^2(\log J + t)}{n}} + \frac{2b(\log J + t)}{3n}$$

with probability at least $1 - e^{-t}$. Replacing a function by its net center changes $(P - P_n)g$ by at most $2/n$, which is absorbed in $b(d_n + t)/n$.

C TRANSFORMER APPROXIMATION OF REGULARIZED TREE CONDITIONALS

In this appendix, we construct a sparse Transformer approximating the regularized logits defined in Appendix B.5. We first consider a single tree in Appendix C.1, and then combine the component constructions in Appendix C.2. We recall the message identities in Appendix C.3, and give the height-dependent single-tree and mixture constructions in Appendices C.4 and C.5, respectively.

Throughout this section, let $L, V \geq 2$, $J \geq 1$, $0 < \delta < 1$, and $0 < \varepsilon \leq 1/2$. Write

$$\rho_\delta := \delta/V, \quad G_\delta := \log(V/\delta), \quad H := (L+1)G_\delta + \log(64LJ/\varepsilon), \quad b_* := \log(eLJVH/\varepsilon).$$

For $c \in \overline{\mathbb{V}}^L$, let $W(c) := \{i : c_i \neq \mathtt{M}\}$ and write $q(c) := \mathbb{P}_q[X_{W(c)} = c_{W(c)}]$. We use the parameter class Θ of Section 2.2 and its clipped logit class $\mathcal{F}(\Theta)$ from Appendix B.3. Furthermore, when constructing approximations, we denote LSE as the log-sum-exp operator defined in Lemma 18.

C.1 SINGLE-TREE CONSTRUCTION BY PARALLEL CONTRACTION AND RECONSTRUCTION

Lemma 9 (Single-tree parallel construction)**.** *Let q be a J-state hidden Markov tree model on $[L]$, with root o, root distribution ν, transition tables A_i, and emission tables B_i satisfying $B_i(v,z) \geq \rho_\delta$. There exists a parameter class $\Theta = \Theta(D, N, M, D_{\mathrm{ff}}, B, S; \sqrt{3}, G_\delta)$ containing a predictor such that*

$$\sup_{c:\, W(c)\neq[L]} \max_{i\notin W(c),\, v\in\mathbb{V}} |F_{\theta,i,v}(c) - \log q_i(v \mid c)| \leq \varepsilon. \tag{63}$$

The architectural budgets can be chosen as

$$\begin{aligned} &N \leq Cb_*^3, \qquad M \leq 2, \qquad \log B \leq Cb_*^2, \\ &D \leq C\big[L + V + (J^3 + JV)b_*^3\big], \qquad D_{\mathrm{ff}} \leq C\big[LV + (J^3 + JV)b_*^3\big], \\ &S \leq C\big[L(J^2 + JV) + J^3 + JV\big]b_*^{10}. \end{aligned} \tag{64}$$

Before the output embedding, the construction also contains a coordinate $\widehat{h}_i(c)$ at every row satisfying

$$\sup_{c\in\overline{\mathbb{V}}^L} \max_{i\in[L]} |\widehat{h}_i(c) - \log q(c)| \leq \varepsilon. \tag{65}$$

The budgets are independent of the topology and probability tables. For a fixed topology, the permitted coefficient locations can be fixed independently of the tables.

By Lemma 6, the lemma applies to $q^\delta = P_\delta q$ for any $q \in \mathcal{M}_{J,1}$. In this case, (63) approximates $F^{\delta,q}$ on masked positions. The additional coordinate in (65) will be used later to combine different components.

Proof of Lemma 9. We first eliminate vertices while preserving the marginal law on the remaining tree. Reversing these operations recovers all hidden-state posteriors. We then implement the two passes by transformer blocks and control their errors.

Positive hidden probabilities. Let $u_J = J^{-1}\mathbf{1}$ and define

$$\gamma := \frac{\varepsilon\rho_\delta^L}{64L}, \qquad \nu^\gamma := (1-\gamma)\nu + \gamma u_J, \qquad A_i^\gamma := (1-\gamma)A_i + \gamma u_J\mathbf{1}^\top. \tag{66}$$

Keep the emissions unchanged and denote the resulting observed law by q^γ. A coupling of the L hidden draws gives $\sup_E |q^\gamma(E) - q(E)| \leq L\gamma$, where E ranges over events in the observed sequence space. Both laws assign probability at least $\rho_\delta^{|W|}$ to a context on W, and at least $\rho_\delta^{|W|+1}$ to that context together with one masked token. Thus,

$$\begin{aligned} |\log q_i^\gamma(v \mid c) - \log q_i(v \mid c)| &\leq 2L\gamma\rho_\delta^{-L} = \varepsilon/32, \\ |\log q^\gamma(c) - \log q(c)| &\leq L\gamma\rho_\delta^{-L} = \varepsilon/64. \end{aligned} \tag{67}$$

Here we use $|\log a - \log b| \leq |a-b|/\min\{a,b\}$. Below, the hidden tables refer to the smoothed model, and we suppress their superscript γ. Notice that

$$H = G_\delta + \log(J/\gamma), \qquad (\gamma/J)^L\rho_\delta^L = e^{-LH}.$$

Contraction and the marginal law. Let $\mathcal{V}_0 = [L]$. Initially, set

$$g_i(z,c) := \begin{cases} B_i(c_i, z), & i \in W(c), \\ 1, & i \notin W(c), \end{cases} \qquad \psi_o^{(0)}(z) := \nu(z) g_o(z,c), \quad \psi_i^{(0)}(z) := g_i(z,c) \quad (i \neq o),$$

and $T_i^{(0)} := A_i$ for $i \neq o$. Here $T_i^{(s)}$ denotes an effective edge factor; B_i continues to denote an emission table. Write pa_s for the parent map after round s. We maintain

$$\mathbb{P}_{q^\gamma}[Z_{\mathcal{V}_s} = z_{\mathcal{V}_s}, c] = \prod_{i \in \mathcal{V}_s} \psi_i^{(s)}(z_i) \prod_{i \in \mathcal{V}_s \setminus \{o\}} T_i^{(s)}(z_i, z_{\mathrm{pa}_s(i)}). \tag{68}$$

This holds initially by the hidden-tree factorization. There are two elimination operations. In the formulas below, all factors on the right are taken before the round.

- If v is a nonroot leaf with parent u, define

$$\ell_v(a) := \sum_{b=1}^{J} T_v(b,a) \psi_v(b), \qquad \psi_u(a) \leftarrow \psi_u(a) \ell_v(a),$$

  and remove v.
- If v is a nonroot vertex with parent u and unique child w, define

$$C_v(z,a) := \sum_{b=1}^{J} T_w(z,b) \psi_v(b) T_v(b,a). \tag{69}$$

  Remove v, connect w to u, and replace T_w by C_v.

Both operations sum out Z_v in (68). As in tree contraction Miller & Reif (1985), we apply them to nonadjacent vertices. If several selected leaves share a parent, their factors ℓ_v are multiplied into its unary factor. The eliminated variables have no edges between them, so simultaneous elimination gives exactly the marginal on the surviving tree.

At round s, consider the nonroot vertices in $\mathcal{V}_{s-1}$ with at most one child. Select the larger of their even- and odd-depth sets, breaking ties by choosing even depth, and call it I_s. Put $\mathcal{V}_s := \mathcal{V}_{s-1} \setminus I_s$. For a tree with $n \geq 2$ vertices, if b vertices have degree at least three, then

$$2n - 2 \geq 3b + (n - b), \qquad b \leq (n-2)/2.$$

After excluding the root, at least $n/2$ vertices have at most one child. One depth parity therefore contains at least $n/4$ eligible vertices, and vertices of the same parity are nonadjacent. Consequently,

$$|I_s| \geq |\mathcal{V}_{s-1}|/4, \qquad |\mathcal{V}_s| \leq \tfrac{3}{4}|\mathcal{V}_{s-1}|, \qquad s_0 \leq 1 + \lceil \log_{4/3} L \rceil, \tag{70}$$

where $\mathcal{V}_{s_0} = \{o\}$. The schedule depends only on the topology.

Reconstruction of the posteriors. When removing $v \in I_s$, save $\psi_v^{(s-1)}$ and $T_v^{(s-1)}$, together with ℓ_v if v is a leaf, or $T_w^{(s-1)}$ and C_v if it has one child w. Each row is removed once, so these saved factors require $O(J^2)$ coordinates per row.

At the end of contraction, (68) gives

$$\psi_o^{(s_0)}(a) = \mathbb{P}_{q^\gamma}[Z_o = a, c], \qquad q^\gamma(c) = \sum_a \psi_o^{(s_0)}(a), \qquad p_o(a) := \frac{\psi_o^{(s_0)}(a)}{q^\gamma(c)}.$$

Suppose that, before reversing round s, we know

$$p_u(a) = \mathbb{P}_{q^\gamma}[Z_u = a \mid c] \quad (\text{for } u \in \mathcal{V}_s),$$
$$P_u^{(s)}(z,a) = \mathbb{P}_{q^\gamma}[Z_u = z, Z_{\mathrm{pa}_s(u)} = a \mid c] \quad (\text{for } u \in \mathcal{V}_s \setminus \{o\}).$$

We recover the node and edge posteriors on $\mathcal{V}_{s-1}$ as follows.

- If v was a leaf with parent u, then

$$P_v^{(s-1)}(b,a) = p_u(a)\frac{\psi_v^{(s-1)}(b)T_v^{(s-1)}(b,a)}{\ell_v(a)}, \qquad p_v(b) = \sum_a P_v^{(s-1)}(b,a). \tag{71}$$

- If v had parent u and child w, form

$$\Gamma_v(z,b,a) := P_w^{(s)}(z,a)\frac{T_w^{(s-1)}(z,b)\psi_v^{(s-1)}(b)T_v^{(s-1)}(b,a)}{C_v(z,a)}. \tag{72}$$

Then

$$P_v^{(s-1)}(b,a) = \sum_z \Gamma_v(z,b,a), \quad P_w^{(s-1)}(z,b) = \sum_a \Gamma_v(z,b,a),$$
$$p_v(b) = \sum_a P_v^{(s-1)}(b,a).$$

Indeed, conditionally on the surviving neighbors, the factors involving Z_v in (68) are precisely the numerators in these kernels; summing over Z_v gives their denominators. Thus Γ_v is the joint posterior of (Z_w, Z_v, Z_u) given c. Nonadjacency ensures that all required neighbors belong to $\mathcal{V}_s$. The candidates use only the saved factors and the posteriors available before the reverse round. We form all candidates before replacing the affected edges, and retain the existing node and unchanged edge posteriors. This proves the induction and recovers every p_i after s_0 reverse rounds.

Logarithmic coordinates and initialization. All live factors are at most one. Indeed, unary factors are initially at most one, and the initial edge matrices are column stochastic. Leaf elimination preserves this bound, while

$$\sum_z C_v(z,a) = \sum_b \psi_v(b)T_v(b,a)\sum_z T_w(z,b) \le 1$$

shows that effective edge matrices remain column substochastic. Every factor is a sum of products of distinct original factors. Retaining one compatible assignment bounds it below by e^{-LH}. The same lower bound holds for node, edge, and three-node posteriors: retain one compatible full hidden assignment and divide by $q^\gamma(c) \le 1$. It also holds for each reconstruction kernel, whose numerator uses disjoint original factors and whose denominator is at most one. We therefore clip their logarithms to $[-LH, 0]$; clipping cannot increase error relative to the exact quantities.

Use separate orthonormal coordinate vectors p_i for positions, e_v for tokens and the mask, and b for a constant. Choose the trainable embeddings as

$$E_{v,\cdot} = e_v^\top + b^\top, \qquad P_{i,\cdot} = \mathsf{p}_i^\top, \qquad \mathsf{Embed}_{E,P}(c)_{i,\cdot} = \mathsf{p}_i^\top + e_{c_i}^\top + b^\top.$$

Thus $R_{\mathrm{embed}} = \sqrt{3}$ and these matrices use $2(V+1)+L$ nonzeros. Auxiliary coordinates are initially zero. Besides the JV emission entries, allocate $O(J^2)$ coordinates for the live and saved factors, node and edge log posteriors, and receive and candidate quantities. The temporary J^3 triple entries and the arithmetic workspace are reused across rounds.

For completeness, the position-dependent log tables can be loaded exactly by one attention head. Let C_i be the vector containing the entries of $\log A_i$, $\log B_i$, and $\mathbf{1}\{i = o\}\log\nu$, in separate auxiliary coordinates; put zero in the unused root edge entries. Set

$$Q = \sum_j p_j p_j^\top + bb^\top, \qquad K = 2\log 3 \sum_j p_j p_j^\top - \log 3\, bb^\top,$$
$$V_{\mathrm{tab}} = 2L\sum_j p_j\left(C_j - \frac{1}{L+2}\sum_\ell C_\ell\right)^\top.$$

The sigmoid attention weights are $3/4$ on the diagonal and $1/4$ elsewhere. Hence the attention output, including the factor $1/L$, equals $C_i^\top$ at row i. A rowwise FFN then writes

$$\log g_i(z,c) = -\sum_{v\in\mathbb{V}} \mathrm{ReLU}\left(-\log B_i(v,z) + G_\delta \mathbf{1}\{c_i = v\} - G_\delta\right).$$

Adding $\log\nu$ at the root initializes $\log\psi_i^{(0)}$. These operations use $O(L(J^2+JV))$ nonzeros and FFN width $O(JV)$, including the input embeddings.

Routing and rowwise operations. For a prescribed relation $\mathcal{E} \subseteq [L]^2$, let $\mathcal{J}(\mathcal{E}) := \{j : (i,j) \in \mathcal{E} \text{ for some } i\}$. The pair (i,j) means that row i receives from row j. For source coordinates f_ℓ and initially zero receive coordinates r_ℓ, set

$$Q_{\mathcal{E}} = \sum_{(i,j)\in\mathcal{E}} \mathsf{p}_i p_j^\top + bb^\top, \quad K_{\mathcal{E}} = 2\lambda \sum_{j\in\mathcal{J}(\mathcal{E})} \mathsf{p}_j p_j^\top - \lambda bb^\top, \quad V_{\mathcal{E}} = L\sum_{\ell} f_\ell r_\ell^\top. \tag{73}$$

The scores equal λ on $\mathcal{E}$ and $-\lambda$ elsewhere. If source entries lie in $[-LH, 0]$, each received entry differs from its prescribed sum by at most $L^2He^{-\lambda}$. This includes leakage from rows outside $\mathcal{J}(\mathcal{E})$.

A positional affine map writes $d = \mathbf{1}\{i \in I\}$ for any prescribed active set I. For $|x| \le LH$ and $d \in \{0,1\}$, the following identities implement selection, clipping, and reset by residual FFNs:

$$\begin{aligned} d \cdot x &= \mathrm{ReLU}(x) - \mathrm{ReLU}(x - LHd) - \mathrm{ReLU}(-x) + \mathrm{ReLU}(-x - LHd), \\ \mathrm{Clip}_{[-LH,0]}(x) &= -LH + \mathrm{ReLU}(x + LH) - \mathrm{ReLU}(x), \\ -x &= -\mathrm{ReLU}(x) + \mathrm{ReLU}(-x). \end{aligned}$$

To update a stored value, clip the candidate first and apply the first identity to the candidate minus the old value. Inactive rows are then preserved exactly. The constant coordinate supplies affine offsets. For m scalar coordinates, these operations have width and sparsity $O(m)$; the active-set indicator uses $O(|I| + 1)$ nonzeros.

Let

$$\xi := \frac{\varepsilon}{2^{14}L(s_0+1)}, \qquad t := \log(eJ/\xi), \qquad \lambda := \log(L^2H/\xi). \tag{74}$$

By Lemmas 18 and 13, a d-input LSE can be approximated to accuracy ξ using residual rowwise FFNs with depth Ct^2, width Cdt^3, sparsity Cdt^8, and log Frobenius norm at most Ct^2, for the arities used below. Parallel copies use disjoint workspace coordinates. We reset the receive, candidate, indicator, and workspace coordinates after every stage, and leave the persistent coordinates unchanged during the arithmetic.

In a contraction round, rowwise FFNs compute

$$\log \ell_v(a) = \mathrm{LSE}_b\big(\log T_v(b,a) + \log \psi_v(b)\big),$$

$$\log C_v(z,a) = \mathrm{LSE}_b\big(\log T_w(z,b) + \log \psi_v(b) + \log T_v(b,a)\big).$$

For a selected one-child vertex, attention first copies its child's edge table. After computing the candidates, attention aggregates the selected leaf log messages at their surviving parents and sends each replacement edge table to its surviving child. The updates affect only the corresponding rows. The eliminated rows retain their saved factors throughout all later contraction rounds.

At the root, one LSE gives the log evidence, and subtraction initializes the log posterior. In a reverse round, attention copies the parent's node posterior to a removed leaf, or the surviving child's edge posterior to a removed one-child vertex. Affine operations form the log kernels in (71)–(72). Further LSEs marginalize the joint posteriors. The restored child-edge table is sent to its child only after all candidates have been computed. Each round uses a constant number of routing and arithmetic groups. The largest group consists of $O(J^2)$ LSEs with J inputs, and therefore has width $O(J^3t^3)$ and sparsity $O(J^3t^8)$. These are shared rowwise FFNs, so their parameter count has no additional factor L.

Error control. We count the original factors represented by each live factor. Assign weight two to the initial root unary and weight one to every other initial unary and edge. The total weight is $2L$. Elimination combines disjoint collections of original factors. Induction gives a maximum log error of $4(w-1)\xi$ for a live factor of weight w. To verify the induction, a leaf message adds one LSE error and its aggregation at the parent adds one routing error; contraction of a one-child vertex adds at most three local errors, from copying, LSE, and sending. The sums of the input error bounds leave at least this amount in $4(w-1)\xi$. A unary factor is updated only if it has a selected leaf. Clipping is nonexpansive relative to the exact target. Consequently,

$$\text{live-factor error} \le 8L\xi, \qquad \text{saved-factor error} \le 10L\xi.$$

The second bound includes a copied old child table and the stored elimination denominator. The root log-evidence error is at most $(8L+1)\xi$, and its log-posterior error is at most $17L\xi$.

Let e_s bound all node and edge log-posterior errors on $\mathcal{V}_s$ before reversing round s. A leaf kernel uses three saved quantities, and a one-child kernel uses four. Each reconstructed joint posterior uses only one pre-round posterior. Routing and marginalization contribute at most 4ξ along each output calculation. Therefore,

$$e_{s_0} \le 17L\xi, \qquad e_{s-1} \le e_s + 50L\xi, \qquad e_0 \le 70L(s_0+1)\xi.$$

For a masked position, conditional independence of its emission gives

$$\log q_i^{\gamma}(v \mid c) = \mathrm{LSE}_z\big(\log p_i(z) + \log B_i(v,z)\big).$$

Using V parallel LSEs adds at most ξ, so the error is at most $71L(s_0+1)\xi$. Clip the result to $[-G_\delta, 0]$ in output coordinates o_v and take $W_{\mathrm{out}} = (o_v)_{v\in\mathbb{V}}$. This matrix has V nonzeros and Frobenius norm $\sqrt{V}$. Combining the bound with (67) proves (63).

Finally, retain the root log evidence, clip it to $[-LH, 0]$, and broadcast it to every row using (73). Its error relative to $\log q^{\gamma}(c)$ is at most $(8L+2)\xi$. The second inequality in (67) proves (65), including fully observed contexts.

Architectural budgets. Initialization uses $O(L(J^2+JV))$ nonzeros. Each forward or reverse round uses a constant number of routing relations with $O(L)$ pairs, together with the rowwise arithmetic described above. The final emissions use $O(JVt^8)$ nonzeros. Thus,

$$\begin{aligned}
N &\le C(s_0+1)t^2, \qquad M \le 2,\\
D &\le C\big[L+V+J^2+JV+(J^3+JV)t^3\big],\\
D_{\mathrm{ff}} &\le C\big[LV+J^2+JV+(J^3+JV)t^3\big],\\
S &\le C\big\{L(J^2+JV)+(s_0+1)(L+J^3t^8)+JVt^8\big\}.
\end{aligned}$$

The input embeddings, output embedding, gates, and evidence broadcast are included in these bounds. Each row is eliminated once, so saving the factors introduces no factor s_0 in the width. Since $s_0, t \le Cb_*$ and $J^2 \le J^3$, these bounds imply (64). Table loading, routing, and gates have coefficients polynomial in (L, J, V, H, λ), whereas the LSE coefficients are bounded by $\exp(Ct^2)$. Converting to matrix Frobenius norms adds only polynomial dimension factors, giving

$$\log B \le C\big[t^2 + \log(eL^2JVH)\big] \le Cb_*^2.$$

For each topology, the schedule fixes the routing and selection locations. All table and arithmetic locations can be allocated in advance, independently of their values. This proves the lemma.

C.2 Mixtures of trees and the approximation bound

In this section, we combine the single-tree constructions to approximate $F^{\delta,q}$ for $q \in \mathcal{M}_{J,R}$. The component context likelihoods determine their posterior mixture weights and must be retained along with their conditional logits.

Proposition 3 (Uniform approximation of regularized mixture logits)**.** *Fix $R \ge 1$ and set*

$$H_R := (L+1)G_\delta + \log(64LJR/\varepsilon), \qquad b_R := \log(eLJVRH_R/\varepsilon).$$

There exists one parameter class $\Theta = \Theta(D, N, M, D_{\mathrm{ff}}, B, S; \sqrt{3}, G_\delta)$ such that

$$\sup_{q\in\mathcal{M}_{J,R}} d_{\infty,\mathrm{mask}}(F^{\delta,q}, \mathcal{F}(\Theta)) \le \varepsilon, \tag{75}$$

with

$$\begin{aligned}
&N \le Cb_R^3, \qquad M \le 2R, \qquad \log B \le Cb_R^2,\\
&D \le C\big[L+V+R(J^3+JV)b_R^3\big], \qquad D_{\mathrm{ff}} \le C\big[LV+R(J^3+JV)b_R^3\big],\\
&S \le CR\big[L(J^2+JV)+J^3+JV\big]b_R^{10}.
\end{aligned} \tag{76}$$

For mixtures sharing a common rooted topology, two heads suffice. The same bounds hold when the component topologies are restricted to any prescribed family of rooted trees on $[L]$.

Proof of Proposition 3. Fix $q = \sum_{r=1}^{R} \pi_r q_r \in \mathcal{M}_{J,R}$, allowing zero weights. By Lemma 6, $q^\delta = \sum_r \pi_r q_r^\delta$, and each component has emissions $B_{i,r}^\delta(v,z) = (1-\delta)B_{i,r}(v,z) + \delta/V$.

Component constructions and mixture weights. Apply Lemma 9 to each q_r^δ with accuracy $a := \varepsilon/64$. Retain its conditional logits and log evidence as internal coordinates, denoted by $\widehat{f}_{i,r,v}$ and $\widehat{h}_{i,r}$. They satisfy

$$|\widehat{f}_{i,r,v}(c) - \log q_{r,i}^\delta(v \mid c)| \le a, \qquad |\widehat{h}_{i,r}(c) - \log q_r^\delta(c)| \le a.$$

The first bound is for masked positions, and the second holds for all contexts. The hidden-table perturbation is already included in these bounds. The component constructions occupy disjoint auxiliary coordinates, share the input embedding, and run in parallel. Shorter contraction schedules are padded with identity rounds.

To make the mixture log weights finite, set

$$\kappa := \frac{\varepsilon \rho_\delta^L}{64}, \qquad \pi_r^\kappa := (1-\kappa)\pi_r + \frac{\kappa}{R}, \qquad \bar{q} := \sum_{r=1}^{R} \pi_r^\kappa q_r^\delta.$$

A coupling of the component label gives $\sup_E |\bar{q}(E) - q^\delta(E)| \le \kappa$. As both distributions have the same emission lower bound, the argument in (67) yields

$$|\log \bar{q}_i(v \mid c) - \log q_i^\delta(v \mid c)| \le 2\kappa \rho_\delta^{-L} = \varepsilon/32. \tag{77}$$

Moreover, $|\log \pi_r^\kappa| \le \log(R/\kappa) \le H_R$.

Combining the components. The exact conditional log probabilities of $\bar{q}$ satisfy

$$\begin{aligned}\log \bar{q}_i(v \mid c) = \mathrm{LSE}_r \left(\log \pi_r^\kappa + \log q_r^\delta(c) + \log q_{r,i}^\delta(v \mid c) \right) \\ - \mathrm{LSE}_r \left(\log \pi_r^\kappa + \log q_r^\delta(c) \right).\end{aligned}$$

Use $V+1$ parallel LSE networks of accuracy a to compute

$$\widehat{h}_i = \Phi_a\big((\log \pi_r^\kappa + \widehat{h}_{i,r})_{r=1}^R\big), \qquad \widehat{v}_{i,v} = \Phi_a\big((\log \pi_r^\kappa + \widehat{h}_{i,r} + \widehat{f}_{i,r,v})_{r=1}^R\big).$$

The mixture weights enter as exact affine offsets. Since LSE is one-Lipschitz in the maximum norm, the two errors are at most $2a$ and $3a$, respectively. Subtracting and clipping to $[-G_\delta, 0]$ gives error at most $5a$ relative to $\log \bar{q}_i(v \mid c)$. Together with (77), this gives

$$|F_{\theta,i,v}(c) - F_{i,v}^{\delta,q}(c)| \le 5\varepsilon/64 + \varepsilon/32 < \varepsilon.$$

This explains why averaging the component conditionals with the prior weights π_r would not suffice: the weights given c are proportional to $\pi_r q_r^\delta(c)$.

A common architectural class. Let s_0 be the largest number of contraction rounds among the components, and take $t = C\log(eLRJ/\varepsilon)$ with a sufficiently large universal constant. The branch depths are maximized, whereas their auxiliary widths and sparsities add. Initialization requires $O(LR(J^2+JV))$ nonzeros. The component constructions and final mixture operations therefore satisfy

$$\begin{aligned}N &\le C(s_0+1)t^2, \\ S &\le CR\big\{L(J^2+JV) + (s_0+1)(L+J^3t^8) + JVt^8\big\}.\end{aligned}$$

The R evidence broadcasts cost $O(RL)$ nonzeros, and the final mixture LSEs cost $O(RVt^8)$; both are absorbed above. Replacing the single-tree accuracy by $\varepsilon/64$ changes its logarithmic factors only by a universal constant, so $s_0, t \le Cb_R$. This proves (76), including the Frobenius bound. At most $2R$ heads suffice. If the topologies agree, use the same contraction schedule and attention matrices, and place the component value maps in disjoint coordinates; two heads then suffice.

For fixed component topologies, the coefficient locations can be chosen independently of their probability tables. The displayed budgets are uniform over all topologies. Padding to these budgets and allowing the sparse supports to vary gives a single parameter class Θ for every $q \in \mathcal{M}_{J,R}$. This proves (75).

Connection with the approximation error. Applying Lemma 7 to (75) yields

$$\varepsilon_{\mathrm{approx}}(\Theta) \le \frac{s_{\max}}{L}\mathsf{Proj}_{J,R}(p^\star) + s_{\max}\log\frac{1}{1-\delta} + 2\varepsilon.$$

This proves Proposition 2.

C.3 Message identities for hidden-tree mixtures

In this section, we state the messages to help derive the target conditionals. They provide the upward and downward updates used in the height-dependent construction.

Lemma 10. *Fix $c \in \overline{\mathbb{V}}^L$ and write $W = W(c)$. Let q be a single hidden Markov tree model with root o, root distribution ν, transitions A_i, and emissions B_i. For each vertex i, let $\mathcal{V}_i$ be the subtree rooted at i, including i. For $S \subseteq [L]$, write c_S for the event $\{X_{W\cap S} = c_{W\cap S}\}$, and define*

$$g_i(z,c) := \begin{cases} B_i(c_i, z), & i \in W, \\ 1, & i \notin W. \end{cases}$$

Define the upward messages by

$$\alpha_i(z,c) := g_i(z,c)\prod_{u\in\mathrm{ch}(i)} m_u(z,c), \qquad i \in [L],$$

$$m_i(a,c) := \sum_{z=1}^{J} A_i(z,a)\alpha_i(z,c), \qquad i \neq o, \tag{78}$$

and the downward messages by $\beta_o(z,c) := \nu(z)$ and

$$\beta_i(z,c) := \sum_{a=1}^{J} A_i(z,a)\beta_{\mathrm{pa}(i)}(a,c)g_{\mathrm{pa}(i)}(a,c)\prod_{u\in\mathrm{ch}(\mathrm{pa}(i))\setminus\{i\}} m_u(a,c), \qquad i \neq o. \tag{79}$$

Here empty products equal one. These messages satisfy

$$\alpha_i(z,c)\beta_i(z,c) = \mathbb{P}_q[Z_i = z, c], \qquad q(c) = \sum_{z=1}^{J}\alpha_i(z,c)\beta_i(z,c). \tag{80}$$

Consequently, whenever $q(c) > 0$, for every $i \notin W$ and $v \in \mathbb{V}$,

$$q_i(v \mid c) = \frac{\sum_{z=1}^{J}\alpha_i(z,c)\beta_i(z,c)B_i(v,z)}{\sum_{z=1}^{J}\alpha_i(z,c)\beta_i(z,c)}. \tag{81}$$

For a mixture $q = \sum_{r=1}^{R}\pi_r q_r \in \mathcal{M}_{J,R}$, compute the messages on each component tree. Then, whenever $q(c) > 0$,

$$q_i(v \mid c) = \frac{\sum_{r=1}^{R}\pi_r\sum_{z=1}^{J}\alpha_{i,r}(z,c)\beta_{i,r}(z,c)B_{i,r}(v,z)}{\sum_{r=1}^{R}\pi_r\sum_{z=1}^{J}\alpha_{i,r}(z,c)\beta_{i,r}(z,c)}. \tag{82}$$

Proof of Lemma 10. Fixing $z_i = z$ in the subtree factorization gives

$$\alpha_i(z,c) = \sum_{(z_u)_{u\in\mathcal{V}_i\setminus\{i\}}}\prod_{u\in\mathcal{V}_i\setminus\{i\}} A_u(z_u, z_{\mathrm{pa}(u)})\prod_{u\in\mathcal{V}_i} g_u(z_u, c).$$

Each summed state ranges over $[J]$. Grouping the sum by child subtrees gives (78), starting with $\alpha_i = g_i$ at leaves. Thus $\alpha_i(z,c)$ is the likelihood of the subtree evidence with its root state fixed to z. This finite-sum definition also applies to states with $\mathbb{P}_q[Z_i = z] = 0$.

We next verify that $\beta_i(z,c) = \mathbb{P}_q[Z_i = z, c_{\mathcal{V}_i^c}]$. It holds at the root since $\mathcal{V}_o^c = \varnothing$. For $i \neq o$, let $j = \mathrm{pa}(i)$. The evidence outside $\mathcal{V}_i$ consists of the evidence outside $\mathcal{V}_j$, the observation at j if

present, and the evidence in the sibling subtrees. The tree factorization therefore gives

$$\mathbb{P}_q[Z_i = z, c_{\mathcal{V}_i^c}] = \sum_{a=1}^{J} A_i(z,a)\mathbb{P}_q[Z_j = a, c_{\mathcal{V}_i^c}]$$
$$= \sum_{a=1}^{J} A_i(z,a)\beta_j(a,c)g_j(a,c) \prod_{u\in \mathrm{ch}(j)\setminus\{i\}} m_u(a,c) = \beta_i(z,c).$$

Factoring the evidence into $\mathcal{V}_i$ and its complement proves (80). At a masked position, the emission at i is absent from the context, so

$$\mathbb{P}_q[X_i = v, c] = \sum_{z=1}^{J} B_i(v,z)\mathbb{P}_q[Z_i = z, c].$$

Dividing by $q(c) > 0$ gives (81). Finally, for the mixture,

$$\mathbb{P}_q[X_i = v, c] = \sum_{r=1}^{R} \pi_r \mathbb{P}_{q_r}[X_i = v, c], \qquad q(c) = \sum_{r=1}^{R} \pi_r q_r(c).$$

Apply the single-tree joint identities to each component and take the ratio to obtain (82). This argument does not require every component to assign positive probability to the context.

C.4 Single-tree construction by upward and downward messages

In this section, we implement the recursions above by processing the tree levels. We use the regularization, embedding, routing, and rowwise arithmetic from Appendix C.1.

Lemma 11 (Single-tree message construction)**.** *Under the assumptions of Lemma 9, suppose the rooted tree has height h. There exists a parameter class $\Theta = \Theta(D, N, M, D_{\mathrm{ff}}, B, S; \sqrt{3}, G_\delta)$ containing a predictor such that*

$$\sup_{c:\, W(c)\neq[L]} \max_{i\notin W(c),\, v\in\mathbb{V}} |F_{\theta,i,v}(c) - \log q_i(v \mid c)| \le \varepsilon, \tag{83}$$

with

$$N \le C(h+1)b_*^2, \qquad M \le 1, \qquad \log B \le Cb_*^2,$$
$$D \le C\big[L + V + (J^2 + JV)b_*^3\big], \qquad D_{\mathrm{ff}} \le C\big[LV + (J^2 + JV)b_*^3\big],$$
$$S \le CL(J^2 + JV)b_*^8. \tag{84}$$

Before the output embedding, the construction also retains a coordinate $\widehat{h}_i(c)$ at every row satisfying

$$\sup_{c\in\overline{\mathbb{V}}^L} \max_{i\in[L]} |\widehat{h}_i(c) - \log q(c)| \le \varepsilon. \tag{85}$$

The budgets depend on the topology only through its height. For a fixed topology, the permitted coefficient locations can be fixed independently of the probability tables.

Proof of Lemma 11. Apply (66) and work with q^γ. The conditional-logit and log-evidence perturbations are bounded by (67). As in Appendix C.1, all relevant log probabilities lie in $[-LH, 0]$. Below, the tables and messages refer to q^γ, and the superscript γ on the hidden tables is suppressed.

Use the embedding and exact table initialization from Appendix C.1 to store $\log A_i$, $\log B_i$, $\log g_i$, and $\log \nu$ at the root. Allocate separate coordinates for the upward and downward log messages, received values, candidates, and LSE workspace. Initialize $\widehat{\beta}_o^{\log} = \log \nu$, with other message and temporary coordinates zero. Here the superscript $\log$ indicates a stored approximation to a log message. Set

$$\eta := \frac{\varepsilon}{128L(h+1)}, \qquad t := \log(eJ/\eta), \qquad \lambda := \log(L^2H/\eta).$$

The attention maps in (73) have error at most η per received coordinate. We use the exact clipping, selection, and reset operations from Appendix C.1, and LSE networks of accuracy η from Lemma 18.

Upward pass. Let h_i be the height of the subtree rooted at i, and process $I_s := \{i : h_i = s\}$ for $s = 0, \dots, h$. Every child of an active vertex has already been processed. Using the relation

$$\mathcal{E}_s^{\uparrow} := \{(i, u) : i \in I_s,\ u \in \mathrm{ch}(i)\},$$

attention aggregates the stored child log messages into initially zero receive coordinates $\widehat{r}_i$. Clip these coordinates to $[-LH, 0]$. Rowwise FFNs then form the candidates

$$\widehat{\alpha}_i^{\log}(z) \leftarrow \mathrm{Clip}_{[-LH,0]}\big(\log g_i(z, c) + \widehat{r}_i(z)\big), \tag{86}$$

$$\widehat{m}_i^{\log}(a) \leftarrow \mathrm{Clip}_{[-LH,0]}\Big(\Phi_\eta\big((\log A_i(z, a) + \widehat{\alpha}_i^{\log}(z))_{z=1}^{J}\big)\Big), \qquad i \neq o. \tag{87}$$

The right-hand side of the second update uses the new candidate for $\widehat{\alpha}_i^{\log}$. We select the active rows only after forming the candidates and preserve the other stored messages. The root needs only the first update. At the end of each stage, reset all receive, candidate, indicator, and workspace coordinates.

Let $n_i := |\mathcal{V}_i|$. By (78), induction from the leaves gives

$$\|\widehat{\alpha}_i^{\log} - \log \alpha_i\|_\infty \le \eta + \sum_{u \in \mathrm{ch}(i)} 2n_u\eta = (2n_i - 1)\eta,$$

$$\|\widehat{m}_i^{\log} - \log m_i\|_\infty \le 2n_i\eta. \tag{88}$$

Indeed, routing contributes at most η, and LSE contributes another η. Adding exact local evidence and clipping do not increase these bounds.

Downward pass. Process nonroot vertices by increasing depth. At the beginning of each stage, form

$$\widehat{u}_i(z) := \mathrm{Clip}_{[-LH,0]}\big(\widehat{\alpha}_i^{\log}(z) + \widehat{\beta}_i^{\log}(z)\big). \tag{89}$$

For $D_s := \{i : \mathrm{depth}(i) = s\}$, use

$$\mathcal{E}_s^{\downarrow} := \{(i, \mathrm{pa}(i)) : i \in D_s\}$$

to copy the parent's combined log message. Denote the clipped received vector by $\widehat{r}_i$ and set

$$\widehat{d}_i(a) := \mathrm{Clip}_{[-LH,0]}\big(\widehat{r}_i(a) - \widehat{m}_i^{\log}(a)\big).$$

For $j = \mathrm{pa}(i)$, its exact target is

$$\log \alpha_j(a, c) + \log \beta_j(a, c) - \log m_i(a, c) = \log \beta_j(a, c) + \log g_j(a, c) + \sum_{u \in \mathrm{ch}(j)\setminus\{i\}} \log m_u(a, c)$$

$$= \log \mathbb{P}_{q^\gamma}[Z_j = a, c_{\mathcal{V}_i^c}].$$

Thus subtraction removes the evidence from $\mathcal{V}_i$. The downward recursion (79) is implemented by

$$\widehat{\beta}_i^{\log}(z) \leftarrow \mathrm{Clip}_{[-LH,0]}\Big(\Phi_\eta\big((\log A_i(z, a) + \widehat{d}_i(a))_{a=1}^{J}\big)\Big). \tag{90}$$

As before, form candidates, select the active rows, and reset temporary coordinates. All parent downward messages are complete before they are used. If $e_i := \|\widehat{\beta}_i^{\log} - \log \beta_i\|_\infty$, then

$$e_o = 0, \qquad e_i \le e_{\mathrm{pa}(i)} + (4L + 2)\eta \le \mathrm{depth}(i)(4L + 2)\eta. \tag{91}$$

The two upward-message errors contribute at most $4L\eta$, while routing and LSE contribute at most 2η.

Conditional logits and context likelihood. With both passes complete, form $\widehat{u}_i$ again as in (89). By (80), its target is $\log \mathbb{P}_{q^\gamma}[Z_i = z, c]$. Apply $V + 1$ parallel LSE networks to obtain

$$\widehat{h}_i := \mathrm{Clip}_{[-LH,0]}\Big(\Phi_\eta\big((\widehat{u}_i(z))_{z=1}^{J}\big)\Big), \qquad \widehat{v}_{i,v} := \Phi_\eta\big((\widehat{u}_i(z) + \log B_i(v, z))_{z=1}^{J}\big).$$

Their targets are $\log q^\gamma(c)$ and $\log \mathbb{P}_{q^\gamma}[X_i = v, c]$, respectively; the second identity is needed only when i is masked. Each error is at most

$$\big[2L + h(4L + 2) + 1\big]\eta \le 8L(h + 1)\eta.$$

Clip $\widehat{v}_{i,v} - \widehat{h}_i$ to $[-G_\delta, 0]$, store the result in output coordinates o_v, and use $W_{\mathrm{out}} = (o_v)_{v \in \mathbb{V}}$. This gives conditional-logit error at most $16L(h + 1)\eta = \varepsilon/8$ relative to q^γ. Combining this with (67) proves (83). The retained coordinate $\widehat{h}_i$ has error at most $\varepsilon/16 + \varepsilon/64$ relative to $\log q(c)$, proving (85) for all contexts.

Architectural budgets. Each message stage uses J parallel J-input LSE networks, with depth Ct^2, arithmetic width CJ^2t^3, and arithmetic sparsity CJ^2t^8. For positional routing, every edge and source child is used once in the upward pass. In the downward pass, every edge is used once and each source parent is used at only one depth. The active sets partition the processed vertices, so

$$\sum_s \big(|\mathcal{E}_s| + |\mathcal{J}(\mathcal{E}_s)| + |I_s| + 1\big) \le CL$$

for the upward pass, with I_s replaced by D_s for the downward pass. Here $\mathcal{J}$ is defined before (73). Thus positional routing and indicators cost $O(L)$ over each pass. Value routing and resets use $O(J^2)$ coefficients per stage and are absorbed in the arithmetic count.

Adding initialization, both passes, and the output group yields

$$\begin{aligned} N &\le C(h+1)t^2, \\ S &\le C\big\{L(J^2+JV) + (h+1)J^2t^8 + JVt^8\big\} \le CL(J^2+JV)t^8. \end{aligned}$$

Here we use only $h \le L-1$. Persistent coordinates have dimension $O(L+V+J^2+JV)$, and workspace is reused across stages. The input embeddings and output matrix are counted as in Appendix C.1; retaining $\widehat{h}_i$ adds no routing step. The same coefficient bounds give

$$\log B \le C\big[t^2 + \log(eL^2JVH)\big] \le Cb_*^2.$$

Since $t \le Cb_*$, these estimates give (84). The topology fixes the routing and active sets, and all table and arithmetic locations can be allocated independently of their values.

C.5 Height-dependent construction for mixtures of trees

In this section, we give and prove the height-dependent mixture result of Proposition 4. We replace the component construction in Appendix C.2 by Lemma 11; the mixture-weight perturbation and final LSE operation are unchanged.

Proposition 4 (Height-dependent approximation of regularized mixture logits)**.** *Let $1 \le h \le L-1$, and let $\mathcal{M}^h_{J,R}$ be the subclass of $\mathcal{M}_{J,R}$ whose component trees admit rooted representations of height at most h. With H_R and b_R as in Proposition 3, there exists one class $\Theta = \Theta(D, N, M, D_{\mathrm{ff}}, B, S; \sqrt{3}, G_\delta)$ such that*

$$\sup_{q \in \mathcal{M}^h_{J,R}} d_{\infty,\mathrm{mask}}(F^{\delta,q}, \mathcal{F}(\Theta)) \le \varepsilon, \tag{92}$$

with

$$\begin{aligned} &N \le C(h+1)b_R^2, \qquad M \le R, \qquad \log B \le Cb_R^2, \\ &D \le C\big[L + V + R(J^2+JV)b_R^3\big], \qquad D_{\mathrm{ff}} \le C\big[LV + R(J^2+JV)b_R^3\big], \\ &S \le CRL(J^2+JV)b_R^8. \end{aligned} \tag{93}$$

If all component trees share a common rooted topology, one head suffices. The same bounds hold for any further restriction on the component topologies.

Proof of Proposition 4. Fix a representation $q = \sum_{r=1}^R \pi_r q_r \in \mathcal{M}^h_{J,R}$ with component heights at most h. Regularize the emissions as in Lemma 6 and apply Lemma 11 to each q_r^δ with accuracy $a := \varepsilon/64$. The resulting internal coordinates satisfy

$$|\widehat{f}_{i,r,v}(c) - \log q^\delta_{r,i}(v \mid c)| \le a, \qquad |\widehat{h}_{i,r}(c) - \log q^\delta_r(c)| \le a,$$

where the first bound is for masked positions and the second holds for every context. These estimates already include the hidden-table perturbation. The branches occupy disjoint auxiliary coordinates and share the input embedding. Run the upward stages for all components, followed by their downward stages, with inactive stages padded by identities.

Use the weights π_r^κ and the final $V+1$ LSE networks from Appendix C.2. That calculation uses only the two displayed approximation bounds. The two LSE errors are at most $2a$ and $3a$, and the weight perturbation contributes at most $\varepsilon/32$ by (77). Thus,

$$\sup_{c:\, W(c) \ne [L]} \max_{i \notin W(c),\, v \in \mathbb{V}} |F_{\theta,i,v}(c) - F^{\delta,q}_{i,v}(c)| \le 5a + \varepsilon/32 < \varepsilon.$$

It remains to count the architecture. Take $t = C\log(eLRJ/\varepsilon)$ with a sufficiently large universal constant. Each message stage has depth Ct^2, total arithmetic width CRJ^2t^3, and arithmetic sparsity CRJ^2t^8. The positional routing count in Lemma 11 gives $O(RL)$ nonzeros over each complete pass. Initialization costs $O(LR(J^2+JV))$, the component output groups cost $O(RJVt^8)$, and the final mixture LSEs cost $O(RVt^8)$. Therefore,

$$N \le C(h+1)t^2,$$
$$S \le CR\big\{L(J^2+JV) + (h+1)J^2t^8 + JVt^8\big\} \le CRL(J^2+JV)t^8.$$

Widths and sparsities add across components, while their depths are maximized. All workspace is reused across stages. Since $t \le Cb_R$, the width and Frobenius estimates from Lemma 11 give (93). There is at most one routing head per component at each step. If the rooted topologies coincide, their routing matrices are shared, and the value maps can be combined into one head.

These budgets are uniform over all component trees of height at most h. For fixed topologies, the permitted coefficient locations are independent of the tables. Padding to the displayed budgets and allowing the supports to vary gives one class Θ satisfying (92). This proves the proposition. Finally, Lemma 7 gives the corresponding approximation bound

$$\varepsilon_{\mathrm{approx}}(\Theta) \le \frac{s_{\max}}{L}\mathsf{Proj}_{\mathcal{M}^h_{J,R}}(p^\star) + s_{\max}\log\frac{1}{1-\delta} + 2\varepsilon.$$

The height bound enters the Transformer depth through $h+1$. No logarithmic-height assumption is used.

D Auxiliary neural-network approximation lemmas

We write $\mathsf{NN}(W,d,B,S)$ for feed-forward ReLU networks of hidden width at most W, depth at most d, coefficient magnitude at most B, and at most S nonzero scalar coefficients:

$$\begin{aligned}\mathsf{NN}(W,d,B,S) := \Big\{&\phi(x) = (A_d\,\sigma + b_d)\circ\cdots\circ(A_1 I + b_1)(x)\\ &\big|\ A_i \in \mathbb{R}^{W_{i+1}\times W_i}, b_i \in \mathbb{R}^{W_{i+1}}, \max_i\big\{\|A_i\|_\infty \vee \|b_i\|_\infty\big\} \le B,\\ &\sum_{i=1}^{d}\big(\|A_i\|_0 + \|b_i\|_0\big) \le S\Big\}\end{aligned} \tag{94}$$

Here I denotes the identity map, σ is applied entry-wise using $\mathrm{ReLU}(x) = x \vee 0$, and $W_i \le W$, $\forall i \in [d]$. As a remark, in the auxiliary neural-network class, B bounds individual weights and biases, while in the Transformer class defined in (9), B bounds the Frobenius norm of each parameter matrix. We will convert them when applying results to Transformer structure in Lemma 13.

D.1 Exact operations and residual implementation

Lemma 12 (Exact realization of clipping)**.** *For $a \le b$, define*

$$\mathrm{Clip}_{[a,b]}(x) := \min\{b, \max\{a,x\}\}.$$

This function is represented exactly by a ReLU network in $\mathsf{NN}(2,2,B_{\mathrm{clip}},7)$, where $B_{\mathrm{clip}} := \max\{1,|a|,|b|\}$. Moreover, for all $x, x' \in \mathbb{R}$ and $y \in [a,b]$,

$$|\,\mathrm{Clip}_{[a,b]}(x) - \mathrm{Clip}_{[a,b]}(x')| \le |x-x'|, \quad |\,\mathrm{Clip}_{[a,b]}(x) - y| \le |x-y|.$$

Proof of Lemma 12. The identity

$$\mathrm{Clip}_{[a,b]}(x) = a + \mathrm{ReLU}(x-a) - \mathrm{ReLU}(x-b)$$

gives an exact representation with two hidden units, two affine layers, and at most seven nonzero weights and biases, each of magnitude at most B_{clip}. The clipping function is continuous and piecewise affine with slopes in $\{0,1\}$, hence is 1-Lipschitz. Since $\mathrm{Clip}_{[a,b]}(y) = y$ for $y \in [a,b]$, the second inequality follows from the first.

Lemma 13 (Implementation of ReLU DNNs by Transformer FFN blocks)**.** *Suppose a feed-forward ReLU network has maximum layer dimension W, depth d, coefficient bound $B \geq 1$, and sparsity S. Using a preserved constant coordinate and two reusable buffers, the function can be implemented rowwise by $O(d+1)$ residual FFN blocks as defined in* (6) *with zero attention updates, auxiliary dimension $O(W)$, FFN width $O(W)$, matrix Frobenius norms at most $C(W+1)(B+1)$, and total sparsity*

$$C\big(S + (d+1)W\big).$$

For parallel realization, the sparsities and buffer dimensions add; the squared Frobenius norms add across the branches.

Proof of Lemma 13. Write the ReLU network as

$$z_0 = x, \quad z_\ell = \mathrm{ReLU}(A_\ell z_{\ell-1} + b_\ell) \quad (1 \leq \ell < d), \quad z_d = A_d z_{d-1} + b_d.$$

and pad all layer vectors to dimension W. Set all attention updates to zero, and reserve two groups of W auxiliary coordinates together with a constant coordinate.

Suppose the current row is $h = (z^\top, v^\top, 1)$, where z contains the current NN input and v contains the old values in the destination coordinates. To implement $z \mapsto \mathrm{ReLU}(Az + b)$, choose

$$W_{\mathrm{TF}} = \begin{pmatrix} A^\top & 0 & 0 \\ 0 & I_W & -I_W \\ b^\top & 0 & 0 \end{pmatrix}, \qquad U_{\mathrm{TF}} = \begin{pmatrix} 0 & I_W & 0 \\ 0 & -I_W & 0 \\ 0 & I_W & 0 \end{pmatrix}.$$

Then the residual FFN satisfies

$$\begin{aligned} h + \mathrm{ReLU}(hW_{\mathrm{TF}})U_{\mathrm{TF}} &= \big(z^\top,\, [v + \mathrm{ReLU}(Az+b) - \mathrm{ReLU}(v) + \mathrm{ReLU}(-v)]^\top,\, 1\big) \\ &= \big(z^\top, \mathrm{ReLU}(Az+b)^\top, 1\big). \end{aligned}$$

Thus one block writes the next hidden layer exactly. Alternating the source and destination groups implements successive layers:

$$(z_0, 0) \longmapsto (z_0, z_1) \longmapsto (z_2, z_1) \longmapsto (z_2, z_3) \longmapsto \cdots.$$

For the final affine layer, take $r = Az + b$ and use four groups of hidden activations to produce

$$\Delta v = \mathrm{ReLU}(r) - \mathrm{ReLU}(-r) - \mathrm{ReLU}(v) + \mathrm{ReLU}(-v) = r - v.$$

The residual update therefore gives $v + \Delta v = Az + b$.

To count parameters, let $s_\ell = \|A_\ell\|_0 + \|b_\ell\|_0$. For each hidden layer, the displayed matrices have

$$\|W_{\mathrm{TF},\ell}\|_0 = s_\ell + 2W, \quad \|U_{\mathrm{TF},\ell}\|_0 = 3W,$$

while the final affine layer uses at most $2s_d + 6W$ nonzero coefficients. Including the initial input copy and final coordinate selection yields

$$S_{\mathrm{TF}} \leq C\big(S + (d+1)W\big).$$

The construction uses $O(d+1)$ blocks, auxiliary dimension $O(W)$, and FFN width at most $4W$. Moreover, for each hidden-layer block,

$$\|W_{\mathrm{TF},\ell}\|_F^2 = \|A_\ell\|_F^2 + \|b_\ell\|_2^2 + 2W \leq (W^2 + W)B^2 + 2W, \quad \|U_{\mathrm{TF},\ell}\|_F^2 = 3W.$$

The final block satisfies the same bounds up to constants, so every FFN matrix has Frobenius norm at most $C(W+1)(B+1)$. For parallel networks, use disjoint auxiliary and FFN coordinates; sparsities, auxiliary dimensions, and squared Frobenius norms then add across branches.

D.2 Scalar arithmetic approximations

Lemma 14 (Multiplication on a bounded rectangle)**.** *For $C_0 \geq 1$ and $0 < \epsilon \leq 1/2$, there exists $\phi_{\mathrm{prod}} \in \mathsf{NN}(W, d, B, S)$ such that*

$$\sup_{|x|,|y| \leq C_0} |\phi_{\mathrm{prod}}(x, y) - xy| \leq 4C_0^2\epsilon,$$

where $W = O(1)$, $d = O(\log(e/\epsilon))$, $B = C_0^2$, and $S = O(\log(e/\epsilon))$.

This lemma is a direct corollary of Lemma F.6 in Oko et al. (2023) with two inputs, followed by rescaling.

Lemma 15 (Uniform approximation exponential (Lemma F.12 Oko et al. (2023))). *For $0 < \epsilon \le 1/2$, there exists $\phi_{\exp} \in \mathsf{NN}(W, d, B, S)$ with*

$$W = O(\log \epsilon^{-1}), \quad d = O(\log^2 \epsilon^{-1}), \quad B = \exp\big(O(\log^2 \epsilon^{-1})\big), \quad S = O(\log^2 \epsilon^{-1}),$$

such that for any $x, x' \ge 0$,

$$|e^{-x'} - \phi_{\exp}(x)| \le \epsilon + |x - x'|.$$

Moreover, $|\phi_{\exp}(x)| \le \epsilon$ whenever $x \ge \log 3\epsilon^{-1}$.

Lemma 16 (Uniform approximation of logarithm on the interval). *For $0 < \tau \le 1/2$ and $0 < \epsilon \le 1/2$, there exists $\phi_{\log} \in \mathsf{NN}(W, d, B, S)$ with*

$$W = O\big(\log^2 \epsilon^{-1} \cdot \log \tau^{-1}\big), \qquad d = O\big(\log^2 \epsilon^{-1} \vee \log \tau^{-1}\big),$$
$$B = O\big(\tau^{-1} \vee \log^2 \tau^{-1}\big), \qquad S = O\big((\log \epsilon^{-1} + \log \tau^{-1})^8\big),$$

such that $\sup_{x \in [\tau, 1]} |\phi_{\log}(x) - \log x| \le \epsilon$.

This result follows from Lemma C.9 in Wu & Cai by using $S \lesssim d(W+1)^2$ to control the network sparsity.

D.3 Softmax stability and LSE approximation

Lemma 17 (Global softmax stability). *For any $a, b \in \mathbb{R}^k$,*

$$\|\mathrm{softmax}(a) - \mathrm{softmax}(b)\|_1 \le 2\|a - b\|_\infty.$$

Proof of Lemma 17. Let $h = a - b$, $z(t) = b + th$, and $p(t) = \mathrm{softmax}(z(t))$ for $0 \le t \le 1$. Differentiate it over t gives

$$p_i'(t) = p_i(t)\Big(h_i - \sum_s p_s(t) h_s\Big), \quad \forall i \in [k].$$

Observe that $p_i(t) \in (0, 1)$ together with $\sum_i p_i(t) = 1$, and thus

$$\sum_{i=1}^k |p_i'(t)| \le 2\|h\|_\infty.$$

Integrating over $t \in [0, 1]$ then proves the result.

Lemma 18 (Uniform log-sum-exp approximation). *For $n \ge 1$ and $0 < \xi \le 1/2$, there exists ReLU DNNs $\Phi_\xi \in \mathsf{NN}(W, d, B, S)$ such that*

$$\sup_{x \in \mathbb{R}^n} |\Phi_\xi(x) - \mathrm{LSE}(x)| \le \xi, \quad \textit{where } \mathrm{LSE}(x) := \log \sum_{j=1}^n e^{x_j}.$$

With $t = \log(en/\xi)$ and some universal constant $C > 0$, its netwoek parameters can be chosen as,

$$d \le Ct^2, \quad W \le Cnt^3, \quad S \le Cnt^8, \quad \log B \le Ct^2.$$

It can be implemented by residual rowwise FFNs with the same orders of depth, width, and sparsity, using reusable auxiliary coordinates and a constant input coordinate. The corresponding matrix Frobenius norms also have logarithms at most Ct^2. Moreover, LSE *is 1-Lipschitz with respect to the infinity norm:*

$$|\mathrm{LSE}(x) - \mathrm{LSE}(y)| \le \|x - y\|_\infty. \tag{95}$$

Proof of Lemma 18. For $n = 1$, we just take $\Phi_\xi(x) = x_1$. In the following, assume that $n \ge 2$ and denote

$$\nu := \frac{\xi}{4n}, \quad \eta := \frac{\xi}{2}, \quad \text{and } t := \log(\frac{en}{\xi}).$$

We construct the approximation by composing the exponential and logarithm networks in Lemmas 15 and 16 respectively, together with exact maximum, affine, and clipping operations.

Step 1: exact maximum shift. For $x \in \mathbb{R}^n$, define

$$a(x) := \max_{j\in[n]} x_j, \quad u_j(x) := a(x) - x_j, \quad Z(x) := \sum_{j=1}^{n} e^{-u_j(x)}.$$

Then $u_j(x) \geq 0$, at least one $u_j(x)$ equals zero, and

$$1 \leq Z(x) \leq n, \quad \mathrm{LSE}(x) = \log\Big(e^{a(x)} \sum_{j=1}^{n} e^{x_j - a(x)}\Big) = a(x) + \log Z(x). \tag{96}$$

Apply the following pairwise maximum operation along a balanced binary tree, with comparisons at each level performed in parallel:

$$\max\{v, w\} = \mathrm{ReLU}(v - w) + w.$$

Retaining the original coordinates until the final exact realization of $a(x)$ gives the map

$$\Gamma(x) := \big(a(x), u_1(x), \ldots, u_n(x)\big),$$

realized by ReLU DNNs with depth $O(\log(en))$, width $O(n)$, sparsity $O(n\log(en))$, and coefficient bound $O(1)$. Retaining original coordinates can be realized via ReLU layers using $v = \mathrm{ReLU}(v) - \mathrm{ReLU}(-v)$.

Step 2: approximate the exponential sum. Apply Lemma 15 with accuracy ν to obtain

$$\phi_{\exp,\nu} \in \mathsf{NN}\left(W_{\exp}, d_{\exp}, B_{\exp}, S_{\exp}\right), \quad \text{such that } \sup_{u\geq 0} |\phi_{\exp,\nu}(u) - e^{-u}| \leq \nu,$$

where

$$W_{\exp} \leq Ct, \quad d_{\exp} \leq Ct^2, \quad S_{\exp} \leq Ct^2, \quad \log B_{\exp} \leq Ct^2.$$

Run n copies in parallel and set the approximation of $Z(x)$:

$$\widehat{Z}(x) := \mathrm{Clip}_{[1,n]}\Big(\sum_{j=1}^{n} \phi_{\exp,\nu}(u_j(x))\Big).$$

Clipping is exact by Lemma 12. Since $Z(x) \in [1, n]$, the clipping operator does not introduce additional error. Hence, it holds that,

$$\begin{aligned} |\widehat{Z}(x) - Z(x)| &\leq \Big|\sum_{j=1}^{n} \phi_{\exp,\nu}(u_j(x)) - \sum_{j=1}^{n} e^{-u_j(x)}\Big| \\ &\leq \sum_{j=1}^{n} |\phi_{\exp,\nu}(u_j(x)) - e^{-u_j(x)}| \leq n\nu = \frac{\xi}{4}, \quad \text{with } \widehat{Z}(x) \in [1, n]. \end{aligned} \tag{97}$$

The parallel exponential networks have width $O(nt)$, depth $O(t^2)$, and sparsity $O(nt^2)$. Summation and clipping add $O(n)$ coefficients, of magnitude at most Cn.

Step 3: rescale the logarithm network and compose. Apply Lemma 16 with interval lower bound $\tau = 1/n$ and accuracy η, and hence there exists

$$\phi_{\log,\eta} \in \mathsf{NN}\left(W_{\log}, d_{\log}, B_{\log}, S_{\log}\right), \quad \text{such that } \sup_{v\in[1/n,1]} |\phi_{\log,\eta}(v) - \log v| \leq \eta,$$

where

$$\begin{aligned} W_{\log} &\leq C\log^2(2/\xi)\log n \leq Ct^3, & d_{\log} &\leq C\big(\log^2(2/\xi) \vee \log n\big) \leq Ct^2, \\ B_{\log} &\leq C\big(n \vee \log^2(2/\xi)\big) \leq Cn, & S_{\log} &\leq C\big(\log(2/\xi) + \log n\big)^8 \leq Ct^8. \end{aligned}$$

To transfer the logarithm approximation from $[1/n, 1]$ to $[1, n]$, define the affine maps $A_n(z) := z/n$ and $T_n(v) := v + \log n$. Denote the composition

$$\Lambda_\eta := T_n \circ \phi_{\log,\eta} \circ A_n.$$

It then satisfies

$$\sup_{z\in[1,n]} |\Lambda_\eta(z) - \log z| = \sup_{z\in[1,n]} |\phi_{\log,\eta}(z/n) - \log(z/n)| \le \eta.$$

Consequently, we denote the network construction based on (96) as

$$\Phi_\xi(x) := a(x) + \Lambda_\eta\big(\widehat{Z}(x)\big) = a(x) + \log n + \phi_{\log,\eta}\Big(\frac{1}{n}\mathrm{Clip}_{[1,n]}\Big(\sum_{j=1}^{n}\phi_{\exp,\nu}\big(a(x)-x_j\big)\Big)\Big). \quad (98)$$

Using (96)–(97) and the 1-Lipschitz property of $\log x$ on $[1, n]$, we obtain that, for every $x \in \mathbb{R}^n$,

$$\begin{aligned}|\Phi_\xi(x) - \mathrm{LSE}(x)| &\le \big|\Lambda_\eta\big(\widehat{Z}(x)\big) - \log \widehat{Z}(x)\big| + |\log \widehat{Z}(x) - \log Z(x)| \\ &\le \eta + |\widehat{Z}(x) - Z(x)| \le \frac{\xi}{2} + \frac{\xi}{4} \le \xi.\end{aligned}$$

This gives the construction of uniform approximation of log-sum-exp operator for $x \in \mathbb{R}^n$. Then we summarize its network parameters and its residual realization.

Step 4: network budgets and residual implementation. Compose the preceding networks without multiplying adjacent affine matrices: an intermediate signed vector v can be represented by $(\mathrm{ReLU}(v), \mathrm{ReLU}(-v))$ and reconstructed in the next affine layer. This incurs only constant-factor overhead in the corresponding layer widths and nonzero coefficients; see also Lemmas F.1–F.3 in Oko et al. (2023). The scalar $a(x)$ can be carried through the exponential and logarithm layers and then recovered using two signed channels. Therefore,

$$\begin{aligned} d &\le C\big(\log(en) + d_{\exp} + d_{\log} + 1\big) \le Ct^2, \\ W &\le C\big(n + nW_{\exp} + W_{\log}\big) \le C(nt + t^3) \le Cnt^3, \\ S &\le C\big(n\log(en) + nS_{\exp} + S_{\log} + d_{\exp} + d_{\log} + n\big) \le C(nt^2 + t^8) \le Cnt^8, \\ B &\le C\max\{1, n, B_{\exp}, B_{\log}\} \le \exp\big(O(t^2)\big). \end{aligned}$$

Here sparsities add across the parallel exponential branches; the signed carry channels contribute only $O(d_{\exp} + d_{\log})$ coefficients.

Apply Lemma 13 to this circuit, retaining a constant input coordinate to supply biases. With all attention updates zero, the resulting rowwise residual FFNs satisfy

$$\begin{aligned} N_{\mathrm{FFN}} &\le C(d+1) \le Ct^2, \\ D_{\mathrm{aux}},\ D_{\mathrm{ff}} &\le C(W+n+1) \le Cnt^3, \\ S_{\mathrm{FFN}} &\le C\big(S + (d+1)(W+n+1)\big) \le C(nt^8 + nt^5) \le Cnt^8. \end{aligned}$$

The maximum Frobenius norm B_{F} of the FFN parameter matrices obeys

$$B_{\mathrm{F}} \le C(W+n+1)(B+1) \le \exp\big(O(t^2)\big).$$

These estimates include the reusable workspaces and their reset operations.

Proof of (95). Let $r := \|x - y\|_\infty$. Then

$$y_j - r \le x_j \le y_j + r \quad \text{for every } j \quad \Longrightarrow \quad e^{-r}\sum_{j=1}^{n} e^{y_j} \le \sum_{j=1}^{n} e^{x_j} \le e^{r}\sum_{j=1}^{n} e^{y_j}.$$

Taking logarithms yields $\mathrm{LSE}(y) - r \le \mathrm{LSE}(x) \le \mathrm{LSE}(y) + r$, and thus proves the 1-Lip continuity.