On the Sample Complexity of Next-Token Prediction

Oğuz Kaan Yüksel
EPFL
oguz.yuksel@epfl.ch

Nicolas Flammarion
EPFL
nicolas.flammarion@epfl.ch

Abstract

Next-token prediction with cross-entropy loss is the objective of choice in sequence and language modeling. Despite its widespread use, there is a lack of theoretical analysis regarding the generalization of models trained using this objective. In this work, we provide an analysis of empirical risk minimization for sequential inputs generated by order-k Markov chains. Assuming bounded and Lipschitz logit functions, our results show that in-sample prediction error decays optimally with the number of tokens, whereas out-of-sample error incurs an additional term related to the mixing properties of the Markov chain. These rates depend on the statistical complexity of the hypothesis class and can lead to generalization errors that do not scale exponentially with the order of the Markov chain—unlike classical k-gram estimators. Finally, we discuss the possibility of achieving generalization rates independent of mixing.

1 INTRODUCTION

Large language models (LLMs) such as GPT, LaMDA, and LLaMA (Brown et al., 2020; Thoppilan et al., 2022; Touvron et al., 2023) have achieved unprecedented performance across various tasks, including classical natural language processing tasks such as translation, reasoning, coding, and understanding (Bubeck et al., 2023). These large models are trained on a surprisingly simple task—*next-token prediction*—using massive amounts of data. This paradigm has seen widespread adoption, particularly after the introduction of context-based self-attention (Vaswani et al., 2017) and decoder-only transformers (Radford et al., 2018), which surpasses classical recurrent approaches (Rumelhart et al., 1986; Hochreiter and Schmidhuber, 1997). The context window and self-attention layers in these architectures enable parallelization, which is not attainable with traditional recurrent approaches and is the key to scaling these large language models (Brown et al., 2020). Amid these developments in architecture and scaling, next-token prediction has emerged as the fundamental training objective.

Learning autoregressive next-token prediction models can be conceptualized as learning conditional probabilities $p_t^\star$ from observations of the form

$$\forall t \geq 0\,, \quad x_t \sim p_t^\star\left(\cdot|x_0, \ldots, x_{t-1}\right)\,, \tag{1}$$

where $p_t^\star$ generates a distribution over a vocabulary $\mathcal{D}$ from which the next-token $x_t \in \mathcal{D}$ is sampled. Given their empirical success and the central role of next-token training schemes in recent advances, as also suggested by Malach (2023), understanding how such autoregressive models generalize is of utmost importance.

In practical applications, the functions $p_t^\star$ share a common structure—a transformer architecture with identical parameter weights—and a *context length* $k \geq 1$ through, for example, a *sliding attention window* (Beltagy et al., 2020; Jiang et al., 2023), simplifying Equation (1) into learning a single conditional probability $p^\star$ from observations of the form

$$\forall t \geq 0\,, \quad x_t \sim p^\star\left(\cdot|x_{t-k}, \ldots, x_{t-1}\right)\,. \tag{2}$$

In this case, the learning task reduces to estimating a discrete Markov chain of order k, where the conditional distribution $p^\star$ governs the transitions between tokens.

In this work, we analyze empirical risk minimization using next-token predictors as defined by Equation (2). Learning $p^\star$ then depends on the context length k and the input dimension kd, and ultimately suffers from the curse of dimensionality under no further assumption on $p^\star$. To that end, we consider learning within a class of logit functions with the softmax link σ

$$\mathcal{F}_\Theta \coloneqq \{f_\theta : \theta \in \Theta\}\,, \quad \mathcal{P}_\Theta \coloneqq \{p_\theta = \sigma \circ f_\theta : \theta \in \Theta\}\,,$$

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

that satisfy certain boundness and Lipschitzness assumptions. These assumptions allow us to derive in-sample and out-of-sample prediction rates for the empirical risk minimizer of the next-token prediction loss which generalizes the theory in *i.i.d.* setting to sequential settings.

Our contributions are threefold:

(i) We derive a statistical rate on the *in-sample prediction error* of the empirical risk minimizer for next-token prediction models. This rate does not depend on the mixing properties of the underlying Markov chain, but on the complexity of the hypothesis class Θ, and decays optimally with the number of tokens.

(ii) We derive statistical rates on the *out-of-sample prediction error* that is counterpart to the in-sample rate. These rates necessarily include an additional term related to the mixing properties of the Markov chain.

(iii) We introduce an additional assumption on the generative process, allowing us to derive a generalization rate independent of the mixing properties of the Markov chain.

2 RELATED WORKS

Our work lies at the convergence of four distinct research areas: multinomial logistic regression, nonlinear system identification, Markov chain estimation, and language modeling.

Multinomial logistic regression. For *i.i.d.* observations, the problem simplifies to multinomial logistic regression. Logistic regression is extensively studied in the statistical community, with various approaches available. The logistic loss can be incorporated into the general framework of statistical learning with convex Lipschitz losses (Wainwright, 2019; Chinot et al., 2020), yielding correct dependence on dimension and sample size but with suboptimal exponential dependence on the parameter norm. Utilizing the self-concordance of the loss (Bach, 2010; Ostrovskii and Bach, 2021), this dependence can be improved to polynomial for Gaussian design. Recently, Kuchelmeister and van de Geer (2023) tightened this dependence to logarithmic under the assumption of data generated by a probit model. Similar rates can also be achieved with an alternative improper learning algorithm (Foster et al., 2018). Multinomial logistic regression has received less attention, and we are unaware of works achieving the optimal $O(Kd/n)$ rate for K classes. Abramovich et al. (2021); Levy and Abramovich (2023) focus on sparse multiclass classification but obtain rates scaling with K^2, assuming each class probability is bounded away from zero, resulting in an additional K term in their main result. Loureiro et al. (2021); Tan and Bellec (2024) derive the asymptotic distribution of the maximum-likelihood estimate in multinomial logistic models in the high-dimensional regime, where dimension and sample size are of the same order.

Non-linear system identification. System identification is a well-studied topic in control theory, with recent research emphasizing non-asymptotic statistical guarantees (Tsiamis et al., 2023). Most work has focused on linear systems, for which optimal rates have been established (Simchowitz et al., 2018; Sarkar and Rakhlin, 2019; Yüksel et al., 2024). In contrast, nonlinear system identification has received less attention, with notable exceptions including Oymak (2019); Bahmani and Romberg (2020); Sattar and Oymak (2022); Foster et al. (2020); Kowshik et al. (2021), who study systems of the form $x_{t+1} = \phi(Ax_t) + \varepsilon_t$, where ϕ is a known activation function and ε_t is *i.i.d.* noise. Ziemann et al. (2022) extended these results to the nonparametric setting. However, this framework, which relies on offset Rademacher complexity (Rakhlin et al., 2015; Liang et al., 2015; Kanade et al., 2022), is not applicable to our problem due to state-dependent noise. Additionally, our study relates to Hall et al. (2018), who explore the learning of sparse generalized linear autoregressive processes. The key difference is that in their vector generalized linear autoregressive model, the output coordinates are independent.

Discrete Markov chain estimation. In our setting of discrete observations, our problem is equivalent to the learning of a kth-order discrete Markov chain. The estimation of Markov transition matrices is a well-studied problem (Anderson and Goodman, 1957; Billingsley, 1961; Craig and Sendi, 2002; Falahatgar et al., 2016), with minimax rates provided for the KL divergence by Hao et al. (2018) and the TV-distance by Wolfer and Kontorovich (2019). Notably, the results for the KL divergence are independent of minimal probabilities, but depend on the mixing properties, aligning with our findings. Han et al. (2021) also examine kth-order Markov chains in a different prediction problem, achieving a mixing-independent $\tilde{O}(d^k/n)$ rate. For *i.i.d.* data, the problem simplifies to estimating discrete distributions in KL divergence, a problem well-characterized in the literature (Braess et al., 2002; Paninski, 2004; Kamath et al., 2015).

Neural and n-gram language models. Statistical language models originally utilized n-gram methodologies to compute maximum likelihood estimates with tables of empirical conditional probabilities. How-

ever, with extensive datasets and large context lengths, these tables predominantly remain sparse, and their size scales exponentially with n. Consequently, n-grams have traditionally been limited to small values of $n \leq 5$ (Brants et al., 2007). To address this shortcoming, various smoothing techniques such as backoff—where a smaller context size is used when longer n-grams have a zero count—have been implemented (Jelinek, 1980; Katz, 1987; Liu et al., 2024). To overcome the curse of dimensionality, continuous embeddings and neural networks have been applied to language modeling (Hinton, 1986; Elman, 1990; Schmidhuber and Heil, 1996; Xu and Rudnicky, 2000). Notably, Bengio et al. (2000) first introduced feedforward neural networks to learn shared models as in Equation (2). Mikolov et al. (2010) then used recurrent neural networks to effectively represent the entire context as in Equation (1). Neural language models are now predominantly based on the transformer architecture, as in the GPT framework (Radford et al., 2018). We also note that similar techniques have been broadly used to model general high-dimensional discrete distributions of the form Equation (1) (Bengio and Bengio, 1999; Uria et al., 2016). Additionally, neural language modeling inherently generates word vector embeddings, a phenomenon also explored for general context windows (Collobert and Weston, 2008; Collobert et al., 2011; Mikolov et al., 2013a,b; Bojanowski et al., 2017). Lastly, our setting parallels maximum entropy models (Rosenfeld, 1994; Berger et al., 1996), where multinomial logistic regression is learned from fixed features depending on contexts of arbitrary lengths.

Recently, there has been increasing interest in the intersection of language models and Markov chains, with a focus on representability, learnability, and the loss landscape (Svete and Cotterell, 2024; Svete et al., 2024; Makkuva et al., 2025). More closely related to our work on generalization are the concurrent works of (Zekri et al., 2024) and (Lotfi et al., 2024). The former considers *out-of-sample* generalization error in total variation distance, while the latter focuses on *in-sample* generalization error in negative log likelihood.

3 PRELIMINARIES

For two discrete distributions $p, q \in \Delta^r$, where Δ^r denotes the r-dimensional simplex, the Kullback-Leibler divergence is $\mathcal{D}_{\mathrm{KL}}(p,q) = \sum_{i=1}^{r} p(i) \ln \frac{p(i)}{q(i)}$. Define the total variation as $\|p-q\|_{TV} = \frac{1}{2}\|p-q\|_1$. For any function $f : \mathcal{A} \to \mathcal{B}$, define $f^{\odot k}$ as the function that maps $(a_1, \ldots, a_k)$ to $(f(a_1), \ldots, f(a_k))$. For a matrix $M \in \mathbb{R}^{d_1 \times d_2}$ with singular values $\sigma_1, \ldots, \sigma_{\min\{d_1,d_2\}}$, define its operator norm as $\|M\|_{\mathrm{op}} = \max_\ell |\sigma_\ell|$. Let $\mathbb{1}(\cdot)$ denote the indicator function.

3.1 Data generation process

Let $d, k \in \mathbb{N}^*$ denote the number of tokens and the context length, respectively. Let $\mathcal{D}$ be a dictionary of input tokens of finite size d and let $\mathcal{E}$ represent the canonical embedding or *one-hot* encoding:

$$\mathcal{D} \coloneqq \{\phi, t_1, \ldots, t_d\}, \quad \mathcal{E} : \begin{cases} \mathcal{D} & \to & \mathbb{R}^d \\ \phi & \mapsto & 0 \\ t_i & \mapsto & e_i \end{cases},$$

where ϕ is the special "empty" token and $e_1, \ldots, e_d$ are the canonical basis vectors of $\mathbb{R}^d$.

We consider learning from a logit class $\mathcal{F}_\Theta$, composed of functions parameterized by $\theta \in \Theta$, which take inputs in $\mathbb{R}^{kd}$ and return embeddings in $\mathbb{R}^d$:

$$\mathcal{F}_\Theta \coloneqq \{f_\theta : \mathbb{R}^{kd} \to \mathbb{R}^d \,|\, \theta \in \Theta\}.$$

Assume that all functions f_θ are B-bounded and L-Lipschitz with respect to a norm $\|\cdot\|_p$ on Θ:

Assumption 3.1. *There exists a constant $B \geq 1$ such that $\forall \theta \in \Theta, \nu \in \mathcal{D}^k$*

$$\left\| \left(f_\theta \circ \mathcal{E}^{\odot k}\right)(\nu) \right\|_\infty \leq B.$$

Assumption 3.2. *There exists a constant $L \geq 1$ such that $\forall \theta, \theta' \in \Theta, \nu \in \mathcal{D}^k$:*

$$\| \left((f_\theta - f_{\theta'}) \circ \mathcal{E}^{\odot k}\right)(\nu) \|_2 \leq L \|\theta - \theta'\|_p.$$

For each f_θ and context $\vec{x} = (x_0, \ldots, x_{p-1})$, define the following associated probability $p_\theta(\cdot|\vec{x})$:

$$\forall x \in \mathcal{D}, \quad p_\theta(x|\vec{x}) = \left\langle \sigma\left(f_\theta\left(\mathcal{E}^{\odot k}(\vec{x})\right)\right), \mathcal{E}(x) \right\rangle, \quad (3)$$

where $\sigma : \mathbb{R}^d \to \Delta^d$ denotes the softmax function:

$$\forall i \in [d], \quad \sigma(\nu)_i = \frac{e^{\nu_i}}{\sum_{j \in [d]} e^{\nu_j}}.$$

Thus, $\mathcal{F}_\Theta$ defines a class of autoregressive models $\mathcal{P}_\Theta \coloneqq \{p_\theta : \mathcal{D}^k \to \Delta^d \,|\, \theta \in \Theta\}$ through Equation (3). Furthermore, assume that the data generation process $p^\star$ in Equation (2) belongs to the set $\mathcal{P}_\Theta$:

Assumption 3.3. *There exist a $\theta^\star \in \Theta$ such that the data is generated by $p_{\theta^\star}$, i.e., $p^\star = p_{\theta^\star}$.*

Metric entropy. Our results depend on the complexity of the class Θ, measured by its metric entropy (Wainwright, 2019). Let $\mathcal{N}_\epsilon(\Theta)$ denote an ϵ-net of Θ:

$$\forall \theta \in \Theta, \quad \exists \theta' \in \mathcal{N}_\epsilon(\Theta) : \|\theta' - \theta\|_p \leq \epsilon.$$

Without loss of generality, assume $\mathcal{N}_\epsilon(\Theta)$ is always chosen to have minimal cardinality. The metric entropy of the set Θ is then defined as follows:

Definition 3.4. *The metric entropy is the logarithm of the cardinality of the minimal ϵ-net covering of Θ:*

$$\mathcal{C}_\epsilon(\Theta) \coloneqq \ln\left(\inf_{\mathcal{N}_\epsilon} |\mathcal{N}_\epsilon|\right) .$$

We set $\mathcal{C}(\Theta) \coloneqq \max\{1, \mathcal{C}_1(\Theta)\}$ for convenience.

To relate $\mathcal{C}_\epsilon(\Theta)$ for different ϵ, we assume the following:

Assumption 3.5. *$\mathcal{C}_\epsilon(\Theta)$ is logarithmically decreasing in ϵ, i.e., $\forall 0 < \epsilon \le 1, \mathcal{C}_\epsilon(\Theta) \le \mathcal{C}(\Theta) \ln \frac{e}{\epsilon}$.*

This assumption is mild and verified for many hypothesis classes, including the examples in Section 3.4.

3.2 Empirical risk minimization

We observe N trajectories of length T, generated from $p_{\theta^\star}$ as follows:

$$\forall n \in [N], \quad \forall t \in [T], \quad x_t^{(n)} \sim p_{\theta^\star}(\cdot|\vec{x}_t^{(n)}) ,$$

where $\vec{x}_t^{(n)} = (x_{t-1}^{(n)}, \dots, x_{t-p}^{(n)})$ and $\vec{x}_1^{(i)} \sim \pi_\star$ where $\pi_\star$ is the stationary distribution of $p_{\theta^\star}$. With the exception of Theorem 4.2, we can initialize the trajectories non-stationary, e.g., $x_{-s} = \phi$ for $s \ge 0$.

Define $\hat{\mathcal{L}}(\theta)$ as the training cross-entropy loss associated with the parameter θ:

$$\hat{\mathcal{L}}(\theta) \coloneqq -\frac{1}{NT} \sum_{t,n} \ln p_\theta(x_t^{(n)} \mid \vec{x}_t^{(n)}) ,$$

The maximum likelihood estimate is then given by

$$\hat{\theta} \coloneqq \operatorname{argmin}_{\theta\in\Theta} \hat{\mathcal{L}}(\theta) . \tag{4}$$

To quantify the success of empirical risk minimization, we introduce two metrics. First, define $\tilde{\mathcal{L}}(\theta)$ as the in-sample cross-entropy loss:

$$\tilde{\mathcal{L}}(\theta) \coloneqq \frac{1}{NT} \sum_{t,n} \mathcal{D}_{\mathrm{KL}}(p_{\theta^\star}(\cdot \mid \vec{x}_t^{(n)}) \parallel p_\theta(\cdot \mid \vec{x}_t^{(n)})) . \tag{5}$$

The in-sample prediction error measures the generalization loss of the model on all *seen contexts* and has been studied by, e.g., Foster et al. (2020); Lotfi et al. (2024).

Second, define $\mathcal{L}(\theta)$ as the out-of-sample cross-entropy loss:

$$\mathcal{L}(\theta) \coloneqq \mathbb{E}\left[\mathcal{D}_{\mathrm{KL}}(p_{\theta^\star}(\cdot \mid \vec{x}) \parallel p_\theta(\cdot \mid \vec{x}))\right] , \tag{6}$$

where $\vec{x}$ is a random context uniformly drawn from the fresh sequence $x_1, \dots, x_T$, which is independent of the training data $\{x_t^{(n)}\}_{t,n}$, i.e.,

$$\vec{x} \sim \vec{x}_t , \text{ where } t \sim \mathrm{Unif}([T]) \text{ and } x_{t+1} \sim p_{\theta^\star}(\cdot \mid \vec{x}_t) .$$

The out-of-sample prediction error measures the generalization loss of the model on *unseen contexts* and has been studied by, e.g., Ziemann et al. (2022); Zekri et al. (2024).

3.3 Mixing time

We derive learning guarantees based on the mixing time τ_{mix} of the generative process (Levin and Peres, 2017). To introduce the mixing time, we first introduce some additional notation. Given a context $\vec{x} = (x_1, \dots, x_k)$, consider the sequence generated as

$$\forall i > k , \quad x_i \sim p_{\theta^\star}(\cdot \mid x_{i-k}, x_{i-k+1}, \dots, x_{i-1}) . \tag{7}$$

Let $p_{\theta^\star}^{(t)}(\cdot \mid \vec{x})$ denote the law of $\vec{x}_{(t)} \coloneqq x_t$, produced autoregressively by Equation (7). Additionally, let $p_{\theta^\star}^{(t_1,t_2)}(\cdot \mid \vec{x})$ for $t_1 \le t_2$ denote the joint law of

$$\vec{x}_{(t_1,t_2)} \coloneqq (x_{t_1}, x_{t_1+1}, \dots, x_{t_2}) .$$

Now, we can introduce the mixing time:

Definition 3.6. *Define $\tau_{\mathrm{mix}}(\epsilon, \vec{x}, \vec{y})$ for any two contexts $\vec{x}, \vec{y} \in \mathcal{D}^k$ as:*

$$\tau_{\mathrm{mix}}(\epsilon, \vec{x}, \vec{y}) = \min\{t \ge 0 \,|\, \forall t' \ge t , \\ \|p_{\theta^\star}^{(t'+1,t'+k)}(\cdot \mid \vec{x}) - p_{\theta^\star}^{(t'+1,t'+k)}(\cdot \mid \vec{y})\|_{\mathrm{TV}} \le \epsilon\} .$$

Definition 3.7. *Define $\tau_{\mathrm{mix}}(\epsilon)$ as the mixing time of the generative process:*

$$\tau_{\mathrm{mix}}(\epsilon) \coloneqq \sup_{\vec{x},\vec{y}} \tau_{\mathrm{mix}}(\epsilon, \vec{x}, \vec{y}) .$$

In addition, we consider the following definition:

Definition 3.8. *Define $\tilde{\tau}_{\mathrm{mix}}(\epsilon)$ as:*

$$\tilde{\tau}_{\mathrm{mix}}(\epsilon) \coloneqq \sup_{(\vec{x},\vec{y})\in\mathcal{X}} \tau_{\mathrm{mix}}(\epsilon, \vec{x}, \vec{y}) , \quad \textit{where}$$
$$\mathcal{X} \coloneqq \{(\vec{x}, \vec{y}) \in \mathcal{D}^k \times \mathcal{D}^k \mid \forall 1 \le i < k, \ (\vec{x})_i = (\vec{y})_i\} .$$

Lastly, we set $\tau_{\mathrm{mix}} \coloneqq \tau_{\mathrm{mix}}(e^{-1})$ for brevity.

3.4 Examples

We present three examples of next-token predictors that satisfy the assumptions in Section 3. In what follows, let $\mathcal{M}_B^{d_1\times d_2}$ represent the set of $d_1 \times d_2$ matrices with a bounded operator norm:

$$\mathcal{M}_B^{d_1\times d_2} \coloneqq \{M \in \mathbb{R}^{d_1\times d_2} \mid \|M\|_{\mathrm{op}} \le B\} .$$

Linear next-token predictor. Let $\Theta \coloneqq \{\boldsymbol{A} = (A_1, \dots, A_k) \in \mathcal{M}_B^{d\times kd}\}$ and define $\mathcal{F}_\Theta$ as:

$$\mathcal{F}_\Theta \coloneqq \{f_{\boldsymbol{A}} \mid \boldsymbol{A} \in \Theta\} , \quad \text{where} \quad f_{\boldsymbol{A}}(\vec{x}) \coloneqq \sum_{i=1}^{k} A_i x_{t-i} .$$

Then, the metric entropy of this class is given by

$$\mathcal{C}_\epsilon(\Theta) = \mathcal{O}\left(kd^2 \ln \frac{B}{\epsilon}\right).$$

Notably, the metric entropy scales linearly with the context length k. The cross-entropy loss is convex, so the MLE estimate defined in Equation (4) can be computed using methods such as gradient descent.

The linear next-token predictor is similar to the linear autoregressive models of Malach (2023, Section 2.1.1.), where the softmax is replaced by a maximum. However, our objective differs from Malach (2023) but is complementary: we aim to provide non-asymptotic sample complexity upper-bounds for learning the parameters of the autoregressive next-token predictor, where MLE estimates can be efficiently computed in the case of linear next-token predictors. Malach (2023) provides *universality results*, showing that such models can approximate any function that can be efficiently computed by a Turing machine.

Self-attention next-token predictor. Let $B_Q, B_K, B_V > 0$ be some constants, and let $\mathcal{Q}, \mathcal{K}, \mathcal{V}$ be sets of $d \times d$ matrices, i.e., $\mathcal{Q} \coloneqq \mathcal{M}_{B_Q}^{d\times d}, \mathcal{K} \coloneqq \mathcal{M}_{B_K}^{d\times d}$ and $\mathcal{V} \coloneqq \mathcal{M}_{B_V}^{d\times d}$. For a vector $\vec{\mu} = (\vec{\mu}_1, \ldots, \vec{\mu}_k) \in \mathbb{R}^{kd}$, define the following maps $g_{Q,K} : \mathbb{R}^{kd} \to \mathbb{R}^d$ and $f_{Q,K,V} : \mathbb{R}^{kd} \to \mathbb{R}^d$:

$$g_{Q,K}(\vec{\mu}) \coloneqq \sigma\left(\left(\frac{1}{\sqrt{k}} \sum_{i=1}^{k} \langle Q\vec{\mu}_i, K\vec{\mu}_j\rangle\right)_{j=1}^{d}\right),$$

$$f_{Q,K,V}(\vec{\mu}) \coloneqq V\left(\sum_{i=1}^{k} \left(g_{Q,K}(\vec{\mu})\right)_i \mu_i\right).$$

Then, the logit class is defined as:

$$\mathcal{F}_{\mathcal{Q},\mathcal{K},\mathcal{V}} \coloneqq \{f_{Q,K,V} \mid Q \in \mathcal{Q}, K \in \mathcal{K}, V \in \mathcal{V}\},$$

which leads to a class $\mathcal{P}_{\mathcal{Q},\mathcal{K},\mathcal{V}}$ representing the set of single-layer self-attention models without positional encoding. Fixed positional encodings can be included by shifting $f_{Q,K,V}(\vec{\mu})$ as $f_{Q,K,V}(\vec{\mu} + \vec{\mu}_{\text{pos}})$.

This class verifies Assumption 3.1 with $B = B_V$ and Assumption 3.2 with $L = B_V(B_Q + B_K)$, with respect to the maximum norm:

$$d_{\max}\left((Q_1, K_1, V_1), (Q_2, K_2, V_2)\right) \coloneqq \max\{\|Q_1 - Q_2\|_{\text{op}}, \|K_1 - K_2\|_{\text{op}}, \|V_1 - V_2\|_{\text{op}}\}.$$

The metric entropy of this class is given by

$$\mathcal{C}_\epsilon(\mathcal{F}_{\mathcal{Q},\mathcal{K},\mathcal{V}}) = \mathcal{O}\left(d^2 \ln \frac{B_Q B_K B_V}{\epsilon}\right),$$

We note the metric entropy does not depend on k.

Discrete Markov chains of order k. Let $\Theta \coloneqq \mathcal{M}_D^{d\times d^k}$, where $D > 0$ is a constant, and define $\mathcal{F}_\Theta$ as:

$$\mathcal{F}_\Theta \coloneqq \{f_{\boldsymbol{A}} \mid \boldsymbol{A} \in \Theta\}, \quad \text{where} \quad f_{\boldsymbol{A}}(\vec{x}) \coloneqq \boldsymbol{A}\vec{x}.$$

Thus, $\mathcal{P}_\Theta$ represents the set of discrete Markov chains of order k, where the minimum entry in the transition matrix is bounded away from zero.

The metric entropy of this class is

$$\mathcal{C}_\epsilon(\Theta) = \mathcal{O}\left(d^{k+1} \ln \frac{B}{\epsilon}\right).$$

This class is studied by Hao et al. (2018) for $k = 1$.

4 SAMPLE COMPLEXITY OF NEXT-TOKEN PREDICTION

In this section, we provide learning guarantees for the in-sample and out-of-sample prediction errors defined in Equations (5) and (6).

In-sample error. We begin with the in-sample error in Equation (5), as described in Theorem 4.1:

Theorem 4.1. *Let Assumptions 3.1 to 3.3 and 3.5 hold. For any $0 < \delta < e^{-1}$,*

$$\tilde{\mathcal{L}}(\hat{\theta}) = \mathcal{O}\left(\frac{B\mathcal{C}(\Theta)}{NT} \ln eLNT \ln \frac{1}{\delta}\right),$$

with probability at least $1 - \delta$.

Theorem 4.1 exhibits several noteworthy features.

First, the decay of the error is governed by NT, *the number of total tokens*. To contextualize this result, consider the following two cases: when $p^\star(\cdot \mid \vec{x})$ does not depend on the context $\vec{x}$, which is equivalent to learning a single discrete distribution from *i.i.d.* observations; and when $T = 1$ with uniform initialization $\vec{x}_1$ over the training contexts, which is equivalent to having *i.i.d.* samples in a multi-class classification problem or learning many discrete distributions from *i.i.d.* observations. In both cases, the optimal decay in sample size is $1/NT$. We observe that this optimal dependency is preserved in Theorem 4.1, up to logarithmic terms. This suggests that despite temporal dependencies within the data, the speed of learning still behaves similarly to the *i.i.d* setting.

Second, the error is determined by the metric entropy of the class Θ and does not depend on the mixing time of generative process $p^\star$. This is a crucial feature of Theorem 4.1, as it demonstrates that the error depends on the complexity of the generative process solely through the complexity of the hypothesis class. In particular, the error depends on the order k only through the metric entropy of Θ.

Out-of-sample error. Next, we present the results on the out-of-sample prediction error in Equation (6):

Theorem 4.2. *Let Assumptions 3.1 to 3.3 and 3.5 hold. For any $0 < \delta < e^{-1}$ and $T \gg \tau_{\text{mix}} \ln eT$,*

$$\mathcal{L}(\hat{\theta}) = \mathcal{O}\left(\tau_{\text{mix}} \frac{(B + \ln d)\, \mathcal{C}(\Theta)}{NT} \ln eLNT \ln eNT \ln \frac{1}{\delta}\right),$$

with probability at least $1 - \delta$.

Similar to Theorem 4.1, Theorem 4.2 shows that the error decays with the number of tokens NT, matching the optimal rate in the *i.i.d.* setting, up to logarithmic factors.

Unlike Theorem 4.1, the error in Theorem 4.2 depends on the mixing time τ_{mix} of the generative process. This is expected since the out-of-sample prediction error measures the generalization loss on *unseen contexts*, and the mixing time quantifies how long the generative process takes to explore all contexts. However, for fast-mixing processes, this dependency is mild.

In Example 4.3, we present a simple first-order Markov chain where the dependence on mixing time is necessary for generalization guarantees to hold.

Example 4.3. *Let ϵ be an arbitrarily small constant and let $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \{\phi\}$ be a partition of the dictionary. Consider a first-order Markov chain that obeys the following rules $\forall d_1 \in \mathcal{D}_1, d_2 \in \mathcal{D}_2$:*

$$p^\star(\phi \mid d_1) = p^\star(\phi \mid d_2) = 0\,,$$
$$p^\star(d_2 \mid d_1) = p^\star(d_1 \mid d_2) = \epsilon\,.$$

Moreover, assume that the Markov chain verifies:

$$\sum_{d \in \mathcal{D}_1} p^\star(d \mid \phi) = \sum_{d \in \mathcal{D}_2} p^\star(d \mid \phi) = 1/2\,.$$

The mixing time of the generative process is $\tau_{\text{mix}} = \Omega(1/\epsilon)$. If there is only one sequence of observations, i.e., $N = 1$, with length $T \ll \tau_{\text{mix}}$, then it is likely that the model only explores at most half of the tokens in the dictionary. In such a setting, the generalization error cannot be small without any additional assumptions.

Local mixing. As seen in Example 4.3, the mixing time τ_{mix} can be arbitrarily large, even for first-order Markov chains and generalization is not possible without paying the mixing constant. However, empirical evidence from autoregressive models, such as transformers, suggests that learning is feasible even with slow mixing. This raises the question: what additional structures in the generative process enable learning despite slow mixing? If the generative process allows for faster *local* mixing, a tighter bound can be derived, as presented in Theorem 4.4.

Theorem 4.4. *Let Assumptions 3.1 to 3.3 and 3.5 hold. For any $0 < \delta < e^{-1}$,*

$$\mathcal{L}(\hat{\theta}) \leq \mathcal{O}\left(\frac{B\mathcal{C}(\Theta)}{NT} \ln eLNT \ln \frac{1}{\delta} + \tilde{\tau}_{\text{mix}}\left(\frac{1}{T}\right)(B + \ln d)\sqrt{\frac{\mathcal{C}(\Theta)}{NT} \ln eLNT \ln \frac{1}{\delta}}\right),$$

with probability at least $1 - \delta$.

Note that Theorem 4.4 still applies to the worst-case Markov chains, such as in Example 4.3 where $\tilde{\tau}_{\text{mix}} = \tau_{\text{mix}}$. However, if $\tilde{\tau}_{\text{mix}} \ll NT \ll \tau_{\text{mix}}$, Theorem 4.4 provides a better bound than Theorem 4.2 but with a worse dependence on sample size.

Learning without mixing. Both Theorems 4.2 and 4.4 provide generalization results that require the mixing times of the generative process to be small. However, these requirements are often not suitable for key applications of autoregressive models, such as language modeling, where mixing times can be very large. In order to provide generalization guarantees in such settings, we introduce the following assumption:

Assumption 4.5. *Let $\eta(T) > 0$ be a constant. Assume that for any $(\vec{x}, \vec{y}) \in \mathcal{X}$, there exists a coupling $\gamma(\mu, \nu)$ between the laws of the generative processes, defined by the marginals*

$$\mu \coloneqq p_{\theta^\star}^{(k+1:T)}(\cdot \mid \vec{x}) \quad \textit{and} \quad \nu \coloneqq p_{\theta^\star}^{(k+1:T)}(\cdot \mid \vec{y})\,,$$

such that

$$\mathbb{E}_{\vec{x}_{(k+1:T+k)}, \vec{y}_{(k+1:T+k)} \sim \gamma}\left[\sum_{i=k}^{T+k} \mathbb{1}\left(\vec{x}_{(i)} \neq \vec{y}_{(i)}\right)\right] \leq \eta(T)\,.$$

Assumption 4.5 states that there exists a coupling between the laws of the generative process such that the two sequences that start with $\vec{x}$ and $\vec{x}$, which differ only on the last token, diverge at most η times on average. An intuitive interpretation of Assumption 4.5 is that for any completion of $\vec{x}$, there exists a rephrasing, at most η tokens apart, that completes $\vec{x}$.

Consider the following example in the context of natural language. Let $\vec{x}$ be "The cat is on the table" and $\vec{y}$ be "The cat is on the chair" where $\mathcal{D}$ represents the set of all English words. Then, Assumption 4.5 states that the continuation of the two sentences is at most η tokens apart on average. Intuitively, for majority of the possible continuations, difference between "table" and "chair" would affect only a few tokens.

We state the learning guarantee under Assumption 4.5:

Theorem 4.6. *Let Assumptions 3.1 to 3.3, 3.5 and 4.5 hold. For any* $0 < \delta < e^{-1}$,

$$\mathcal{L}(\hat{\theta}) \leq \mathcal{O}\left(\frac{B\mathcal{C}(\Theta)}{NT}\ln eLNT \ln\frac{1}{\delta} + k\eta(T)\,(B+\ln d)\sqrt{\frac{\mathcal{C}(\Theta)}{NT}\ln eLNT\ln\frac{1}{\delta}}\right),$$

with probability at least $1-\delta$.

Theorem 4.6 shows that under Assumption 4.5, the generalization error is governed by the complexity of the hypothesis class and the number of tokens, and not by the mixing time of the generative process. However, the dependence on the context length k and the coupling constant $\eta(T)$ remains in the bound. If the generative process satisfies Assumption 4.5 with $\eta(T)$ scaling mildly with T, *i.e.*, logarithmically rather than linearly, then Theorem 4.6 offers a generalization guarantee that is strictly sharper than Theorem 4.4, as τ_{mix} and $\tilde{\tau}_{\text{mix}}$ are lower bounded by k.

5 DISCUSSION

In this section, we discuss the rates provided in Section 4 and potential extensions of our results.

Comparison with related work. Closest to our results are the generalization bounds for discrete Markov chains of order k in Hao et al. (2018) and the bounds of Lotfi et al. (2024); Zekri et al. (2024). Theorem 4.2 is analogous to the minimax upper bounds derived by Hao et al. (2018). The advantage of our approach is that the dependency on k is only through the metric entropy of the hypothesis class, whereas in Hao et al. (2018), the dependency on k is exponential. This rate can be matched using Theorem 4.2 by considering the last example in Section 3.4, where the hypothesis class consists of discrete Markov chains of order k.

Theorem 4.1 provides sharper generalization bounds compared to Lotfi et al. (2024), where the authors provide bounds that depend suboptimally on the number of tokens, $1/\sqrt{NT}$. Instead of using the Azuma-Hoeffding inequality, we apply Freedman's inequality, which allows us to obtain sharper concentration inequalities by accounting for the variance of the martingales, resulting in a faster rate of $1/NT$.

Zekri et al. (2024) provide out-of-sample generalization bounds for total variation distance and Kullback-Leibler divergence by utilizing Marton couplings (Paulin, 2015), both with a dependency of $1/\sqrt{NT}$. Theorem 4.2 provides the faster decay rate of $1/NT$ for Kullback-Leibler divergence by exploiting the independent block method (Mohri and Rostamizadeh, 2008) and localization techniques (Wainwright, 2019).

Dependency on the sample size. Theorem 4.2 exhibit the optimal decay rate of $1/NT$, while Theorems 4.4 and 4.6 have a slower decay rate of $\sqrt{1/NT}$. This difference stems from variations in the analyses of the two results. For Theorem 4.2, we employ the independent block method (Mohri and Rostamizadeh, 2008), which allows for localization techniques (Wainwright, 2019), yielding the faster rate. Conversely, Theorems 4.4 and 4.6 rely on the analysis of martingale series, where controlling the variance of martingale differences is not possible, unlike in the independent block method. We see it as an interesting open question to derive the faster variants for Theorems 4.4 and 4.6. This might involve a more refined analysis of the self-normalized martingale series or require additional assumptions on the generative process.

Complex hypothesis classes. In Section 3.4, we provide examples of hypothesis classes that satisfy the assumptions in Section 3. It is possible to generalize our results to more complex hypothesis classes, such as neural networks, by decomposing the complexity of the hypothesis class into the complexity of its constituent parts, where the metric entropy of a Cartesian product can be computed using an ϵ-net that is the Cartesian product of ϵ-nets, leading to a sum over the individual metric entropies. In the case of transformers, this would involve bounding the metric entropy of the self-attention mechanism and the feed-forward networks separately. Regarding the boundedness and Lipschitz assumptions in Assumptions 3.1 and 3.2, they can again be generalized to arbitrary transformer architectures by bounding the operator norm of the weights and the Lipschitz constant of each operation. The dependency on the Lipschitz constant is multiplicative across all Lipchitz constants of the operations in the network. Since our bounds depend logarithmically on the Lipschitz constant, they incur a dependency on the depth of the network. However, the boundedness assumption is more challenging, as the operator norm of the weights in transformers can be unbounded.

Learning without mixing. In Assumption 4.5, we introduce a new assumption that allows for generalization guarantees where mixing dependencies are replaced by a coupling constant and context length. In particular, the assumption requires that the generative process does not diverge significantly after a single token difference, which could be a reasonable assumption for natural language tasks if k is large.

A limitation of this assumption is that it requires couplings for all contexts in the dictionary, including those

that contain ϕ tokens at the beginning. On a semantical level, two texts that are on two different topics might diverge for an arbitrary number of tokens, *i.e.*, it should not be possible to find a coupling that is close for all tokens. So, if k-gram is able to represent these two topics well and the initialization is with empty tokens ϕ, then the assumption cannot be satisfied as there should be a large divergence at the beginning of the generation. However, this is a limitation of the setting, not of the assumption itself. By sampling over the initialization, and asking generalization with respect to the *seen initializations*, we can still provide generalization guarantees under Assumption 4.5. Moreover, the dependency on k arises from each divergence being seen k times autoregressively and can be replaced by the sum of the L_1-Lipschitz constants of class $\mathcal{F}_\Theta$ on each component, if available. We leave these more advanced settings for future research.

Lastly, we remark that it is important to verify Assumption 4.5 in practice. We believe this is a promising direction for future research as Assumption 4.5 or related assumptions could be valuable for establishing theory of learning under sequential data.

6 SKETCH OF PROOFS

In this section, we provide the details of the proofs of theorems in Section 4. By Assumption 3.3, any empirical risk minimizer $\hat{\theta}$ satisfies the following:

$$\hat{\mathcal{L}}(\hat{\theta}) \le \hat{\mathcal{L}}(\theta^\star)\,. \tag{8}$$

A key property of the Kullback-Leibler divergence is that:

$$\begin{aligned}&\tilde{\mathcal{L}}(\theta) = \hat{\mathcal{L}}(\theta) + \hat{Z}(\theta)\,, \quad \text{where} \\ &\hat{Z}(\theta) \coloneqq \frac{1}{NT}\sum_{t,n}\left\langle g_{\theta^\star}(x_t^{(n)}), f_{\theta^\star}(\vec{x}_t^{(n)}) - f_\theta(\vec{x}_t^{(n)})\right\rangle .\end{aligned} \tag{9}$$

Here $\hat{Z}(\theta)$ is the first-order approximation via the gradients with respect to the logits, i.e.,

$$g_{\theta^\star}(x_t^{(n)}) = \nabla_\theta - \ln p_{\theta^\star}(x_t^{(n)} \mid \vec{x}_t^{(n)})\,.$$

Note that $\mathbb{E}[\hat{Z}(\theta)] = 0$ as each of the gradients $g_{\theta^\star}(x_t^{(n)})$ have zero expectation.

Using the self-concordance of the Kullback-Leibler divergence (Bach, 2010) under bounded logits, we can show that Equation (8) implies:

$$\begin{aligned}&\tilde{\mathcal{L}}(\hat{\theta}) \gtrsim \frac{1}{B}\hat{V}(\hat{\theta})\,, \quad \text{where} \\ &\hat{V}(\theta) \coloneqq \frac{1}{NT}\sum_{t,n}\|f_{\theta^\star}(\vec{x}_t^{(n)}) - f_\theta(\vec{x}_t^{(n)})\|^2_{H_{\theta^\star}(\vec{x}_t^{(n)})}\,.\end{aligned} \tag{10}$$

Here $\hat{V}(\theta)$ is the second-order term in the Taylor expansion of $\tilde{\mathcal{L}}(\theta)$ around $\theta_\star$, i.e.,

$$H_{\theta^\star}(\vec{x}_t^{(n)}) \coloneqq \nabla^2_\theta - \ln p_{\theta^\star}(x_t^{(n)} \mid \vec{x}_t^{(n)})\,.$$

Notably, $\hat{V}(\theta)$ is related to variance of the gradients with respect to logits:

$$\hat{V}(\theta) = \frac{1}{NT}\sum_{t,n}\mathbb{E}_{\mathcal{F}_{t,n}}\left\langle g_{\theta^\star}(x_t^{(n)}), f_{\theta^\star}(\vec{x}_t^{(n)}) - f_\theta(\vec{x}_t^{(n)})\right\rangle^2\,,$$

where expectations are with respect to $\mathcal{F}_{t,n}$, the sigma-algebra generated by the first t tokens of the n-th sequence.

Putting Equations (8) to (10) together, we have that

$$\hat{Z}(\hat{\theta}) \ge \tilde{\mathcal{L}}(\hat{\theta}) \gtrsim \frac{1}{B}\hat{V}(\hat{\theta})\,.$$

Note that $\mathbb{E}[\hat{Z}(\theta)] = 0 < \mathbb{E}[\hat{V}(\theta)]$. For a sufficiently large NT, $\hat{Z}(\theta)$ and $\hat{V}(\theta)$ should concentrate around their expectations, and thus the event $\hat{Z}(\theta) \gtrsim \frac{1}{B}\hat{V}(\theta)$ should be low-probability. In particular, $\hat{Z}(\theta)$ is a martingale difference whereas $\hat{V}(\theta)$ is scaled quadratic variation, and we can control its deviations using martingale concentration inequalities.

In-sample prediction error. We control the probability of the event

$$\forall\theta \in \Theta : \gamma\hat{Z}(\theta) \gtrsim \max\{\hat{V}(\theta), \hat{\alpha}\}\,, \tag{11}$$

for arbitrary $\gamma, \hat{\alpha} > 0$, by adopting the techniques of Yüksel et al. (2024). This requires picking α such that whenever $\hat{V}(\theta) \gg \hat{\alpha}$, the event $\gamma\hat{Z}(\theta) \gtrsim \hat{V}(\theta)$ is low-probability and when $\hat{V}(\theta) \ll \hat{\alpha}$, the event $\gamma\hat{Z}(\theta) \gtrsim \hat{\alpha}$ is low-probability.

The intuition is that $\hat{V}(\theta)$ is NT times the variance of the gradients, and $\hat{Z}(\theta)$ is the sum of the gradients, and thus the event $\gamma\hat{Z}(\theta) \gtrsim \hat{V}(\theta)$ is unlikely when variance is large. Whereas when the variance is small, the sum of the gradients concentrate around their mean.

Here, our bounds depend on the metric entropy of the hypothesis class, which is obtained by proving discretized versions of Equation (11) for a finite subset of Θ. We also pay logarithmic factors for the discretization error.

Out-of-sample prediction error. Once the in-sample error is controlled, we can derive the out-of-sample error by using the following decomposition:

$$\mathcal{L}(\hat{\theta}) = \tilde{\mathcal{L}}(\hat{\theta}) + \Big(\mathcal{L}(\hat{\theta}) + \tilde{\mathcal{L}}(\theta)\Big)\,.$$

For all out-of-sample prediction error theorems, we prove a uniform concentration of $\hat{V}(\theta)$ around its expectation $V(\theta)$ in order to control $\mathcal{L}(\theta)$. Each of Theorems 4.2, 4.4 and 4.6 constructs different strategies to achieve this concentration.

Theorem 4.2 uses the independent block method by Mohri and Rostamizadeh (2008) to divide the sequence into almost *i.i.d.* blocks of size that scales with τ_{mix}. This allows for a concentration of $\hat{V}(\theta)$ around its expectation with localization techniques (Wainwright, 2019), which achieve the optimal *i.i.d.* rate in NT. Here, we extend the hypothesis class to a star-shaped set to apply localization techniques.

In contrast, Theorems 4.4 and 4.6 rely on the analysis of $\mathcal{L}(\theta) - \tilde{\mathcal{L}}(\theta)$ as a martingale series:

$$\mathcal{L}(\theta) - \tilde{\mathcal{L}}(\theta) = \sum_{t=1}^{T} \mathbb{E}\left[\tilde{\mathcal{L}}(\theta) \mid \mathcal{F}_t\right] - \mathbb{E}\left[\tilde{\mathcal{L}}(\theta) \mid \mathcal{F}_{t-1}\right] ,$$

where $\mathcal{F}_t$ is the sigma-algebra generated by the first t tokens of each sequence. The key challenge in these proofs is controlling the martingale differences at each step t. Theorem 4.4 leverages local mixing in the generative process to control each martingale difference, while Theorem 4.6 relies on the coupling assumption for this control. Both depends on Assumption 3.1 to control the worst-case deviations.

7 CONCLUSION

In this paper, we derive in-sample and out-of-sample generalization rates for next-token prediction objectives using empirical risk minimization. Unlike traditional k-gram estimation, which requires a number of samples that grows exponentially with k, our results depend on the complexity of the hypothesis class. We show that in-sample error decays at the optimal rate of $1/NT$, independent of the mixing time of the underlying Markov chain. For out-of-sample prediction, the sample complexity depends on the mixing time of the generative process, but maintains the same fast decay rate of $1/NT$. To eliminate mixing time dependencies, we extend out-of-sample results to depend on a *local* notion of mixing time, but this incurs a slowdown, with the error decaying at $1/\sqrt{NT}$ instead of the faster rate of $1/NT$. Lastly, we provide a generalization guarantee without mixing time dependencies by introducing a stronger coupling assumption, where the decay still follows $1/\sqrt{NT}$ and incurs a constant related to the coupling. Crucially, our results rely only on mild Lipschitz and boundedness assumptions. Overall, our theoretical results extend classical generalization theory to the autoregressive setting with discrete data, laying the foundation for further research on generalization.

Acknowledgments

This project was supported by the Swiss National Science Foundation (grant number 212111). We thank Mathieu Even for helpful discussions on the problem formulation phase.

References

Felix Abramovich, Vadim Grinshtein, and Tomer Levy. Multiclass classification by sparse multinomial logistic regression. *IEEE Transactions on Information Theory*, 67(7):4637–4646, 2021.

Theodore W Anderson and Leo A Goodman. Statistical inference about markov chains. *The annals of mathematical statistics*, pages 89–110, 1957.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

Sohail Bahmani and Justin Romberg. Convex programming for estimation in nonlinear recurrent models. *Journal of Machine Learning Research*, 21 (235):1–20, 2020.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

Yoshua Bengio and Samy Bengio. Modeling high-dimensional discrete data with multi-layer neural networks. *Advances in Neural Information Processing Systems*, 12, 1999.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

Adam Berger, Stephen A Della Pietra, and Vincent J Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

Patrick Billingsley. Statistical methods in markov chains. *The annals of mathematical statistics*, pages 12–40, 1961.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

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

Thorsten Brants, Ashok Popat, Peng Xu, Franz Josef Och, and Jeffrey Dean. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, 2007.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust statistical learning with lipschitz and convex loss functions. *Probability Theory and related fields*, 176(3):897–940, 2020.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12:2493–2537, 2011.

Bruce A Craig and Peter P Sendi. Estimation of the transition matrix of a discrete-time markov chain. *Health economics*, 11(1):33–42, 2002.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Moein Falahatgar, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Learning markov distributions: Does estimation trump compression? In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 2689–2693. IEEE, 2016.

Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. Learning nonlinear dynamical systems from a single trajectory. In Alexandre M. Bayen, Ali Jadbabaie, George Pappas, Pablo A. Parrilo, Benjamin Recht, Claire Tomlin, and Melanie Zeilinger, editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 851–861. PMLR, 10–11 Jun 2020. URL `https://proceedings.mlr.press/v120/foster20a.html`.

Dylan J Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Conference on learning theory*, pages 167–208. PMLR, 2018.

Eric C Hall, Garvesh Raskutti, and Rebecca M Willett. Learning high-dimensional generalized linear autoregressive models. *IEEE transactions on information theory*, 65(4):2401–2422, 2018.

Yanjun Han, Soham Jana, and Yihong Wu. Optimal prediction of markov chains with and without spectral gap. *Advances in Neural Information Processing Systems*, 34:11233–11246, 2021.

Yi Hao, Alon Orlitsky, and Venkatadheeraj Pichapati. On learning markov chains. *Advances in Neural Information Processing Systems*, 31, 2018.

Geoffrey E Hinton. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA, 1986.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Frederick Jelinek. Interpolated estimation of markov source parameters from sparse data. In *Proc. Workshop on Pattern Recognition in Practice, 1980*, 1980.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Conference on Learning Theory*, pages 1066–1100. PMLR, 2015.

Varun Kanade, Patrick Rebeschini, and Tomas Vaskevicius. Exponential tail local rademacher complexity risk bounds without the bernstein condition. *arXiv preprint arXiv:2202.11461*, 2022.

Slava Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401, 1987.

Suhas Kowshik, Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Near-optimal offline and streaming algorithms for learning non-linear dynamical systems. *Advances in Neural Information Processing Systems*, 34:8518–8531, 2021.

Felix Kuchelmeister and Sara van de Geer. Finite sample rates for logistic regression with small noise or few samples. *arXiv preprint arXiv:2305.15991*, 2023.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Tomer Levy and Felix Abramovich. Generalization error bounds for multiclass sparse linear classifiers. *Journal of Machine Learning Research*, 24(151):1–35, 2023.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*, 2024.

Sanae Lotfi, Yilun Kuang, Marc Finzi, Brandon Amos, Micah Goldblum, and Andrew G Wilson. Unlocking tokens as data points for generalization bounds on larger language models. *Advances in Neural Information Processing Systems*, 37:9229–9256, 2024.

Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021.

Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Martin Jaggi, Hyeji Kim, and Michael Gastpar. Attention with markov: A curious case of single-layer transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SqZ0KY4qBD.

Eran Malach. Auto-regressive next-token predictors are universal learners, 2023.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013b.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in neural information processing systems*, 21, 2008.

D. Ostrovskii and F. Bach. Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 15(1):326 – 391, 2021. doi: 10.1214/20-EJS1780. URL https://doi.org/10.1214/20-EJS1780.

Samet Oymak. Stochastic gradient descent learns state equations with nonlinear activations. In *conference on Learning Theory*, pages 2551–2579. PMLR, 2019.

Liam Paninski. Variational minimax estimation of discrete distributions under kl loss. *Advances in Neural Information Processing Systems*, 17, 2004.

Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. 2015.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *J. Mach. Learn. Res.*, 16(1):155–186, 2015.

R Rosenfeld. Adaptive statistical language modeling: A maximum entropy approach. *PhD thesis, Carnegie Mellon Univ.*, 1994.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5610–5618. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/sarkar19a.html.

Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *Journal of Machine Learning Research*, 23(140):1–49, 2022.

Jürgen Schmidhuber and Stefan Heil. Sequential neural text compression. *IEEE Transactions on Neural Networks*, 7(1):142–146, 1996.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 439–473. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/simchowitz18a.html.

Anej Svete and Ryan Cotterell. Transformers can represent n-gram language models. *arXiv preprint arXiv:2404.14994*, 2024.

Anej Svete, Nadav Borenstein, Mike Zhou, Isabelle Augenstein, and Ryan Cotterell. Can transformers learn n-gram language models? *arXiv preprint arXiv:2410.03001*, 2024.

Kai Tan and Pierre C Bellec. Multinomial logistic regression: Asymptotic normality on null covariates in high-dimensions. *Advances in Neural Information Processing Systems*, 36, 2024.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. Lamda: Language models for dialog applications, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J Pappas. Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023.

Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. *Journal of Machine Learning Research*, 17(205):1–37, 2016.

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. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Geoffrey Wolfer and Aryeh Kontorovich. Minimax learning of ergodic markov chains. In *Algorithmic Learning Theory*, pages 904–930. PMLR, 2019.

Wei Xu and Alexander Rudnicky. Can artificial neural networks learn language models? 2000.

Oğuz Kaan Yüksel, Mathieu Even, and Nicolas Flammarion. Long-context linear system identification, 2024. URL `https://arxiv.org/abs/2410.05690`.

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.

Ingvar M Ziemann, Henrik Sandberg, and Nikolai Matni. Single trajectory nonparametric learning of nonlinear dynamics. In *conference on Learning Theory*, pages 3333–3364. PMLR, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Section 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable] The setting is theoretical and does not involve any algorithm.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Sections 3 and 4.
   (b) Complete proofs of all theoretical results. [Yes] See Appendix.
   (c) Clear explanations of any assumptions. [Yes] See Sections 3 and 4.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

ORGANIZATION OF THE APPENDIX

The supplemantary material is organized as follows,

- In Appendix A, we provide the preliminary tools needed for our analyses: Kullback-Leibler divergence, self-concordance, the Azuma-Hoeffding inequality and Freedman's inequality.
- In Appendix B, we prove Theorem 4.1.
- In Appendix C, we prove Theorem 4.2.
- In Appendix D, we prove Theorems 4.4 and 4.6.

A PRELIMINARY TOOLS

A.1 Kullback-Leibler divergence

We aim to show a useful property of Kullback-Leibler divergence with a softmax link in Proposition A.1. Later, we relate the Hessian to the square of its gradients in Proposition A.2. Lastly, we show a worst-case bound on the loss in Proposition A.3.

Let $\tilde{\mathcal{D}}_{\mathrm{KL}}(\cdot \parallel \cdot) : \Delta^d \times \mathbb{R}^d \to \mathbb{R}$ be the following function,

$$\tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel q) \coloneqq \mathcal{D}_{\mathrm{KL}}(p \parallel \sigma(q))\,, \text{ for } p \in \Delta^d, q \in \mathbb{R}^d\,, \tag{12}$$

where $\mathcal{D}_{\mathrm{KL}}$ is the usual Kullback-Leibler divergence, i.e.,

$$\mathcal{D}_{\mathrm{KL}}(p_1 \parallel p_2) \coloneqq \sum_{i\in[d]} (p_1)_i \ln \frac{(p_1)_i}{(p_2)_i}\,, \text{ for } p_1 \in \Delta^d, p_2 \in \Delta^d\,.$$

For $p \in \Delta^d$, the gradient and the Hessian of the function $\tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel \cdot)$ are given by the following computations:

$$\frac{\partial \tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel q)}{\partial q} = -\sum_{i\in[d]} \frac{p_i}{\sigma(q)_i} \frac{\partial \sigma(q)_i}{\partial q} = -\sum_{i\in[d]} \frac{p_i}{\sigma(q)_i} \sigma(q)_i \left(e_i - \sigma(q)\right) = \sigma(q) - p\,, \tag{13}$$

$$\frac{\partial^2 \tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel q)}{\partial q \partial q^\top} = \Delta\sigma(q) \coloneqq \mathrm{diag}(\sigma(q)) - \sigma(q)\sigma(q)^\top\,. \tag{14}$$

Proposition A.1. *Let p be a probability distribution in Δ^d and q, q' be vectors in $\mathbb{R}^d$. Then,*

$$\tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel q) - \tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel q') = \left\langle \partial_{q'} \tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel q'), q - q' \right\rangle + \tilde{\mathcal{D}}_{\mathrm{KL}}(\sigma(q') \parallel q)\,.$$

Proof. By Equation (12),

$$\begin{aligned}\tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel q) - \tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel q') &= \sum_{i\in[d]} p_i \ln \frac{\sigma(q')_i}{\sigma(q)_i} \\ &= \langle p, q' - q \rangle + \ln \frac{\sum_{j\in[d]} e^{q'_j}}{\sum_{j\in[d]} e^{q_j}}\,.\end{aligned}$$

And, by Equations (12) and (13),

$$\begin{aligned}\left\langle \partial_{q'} \tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel q'), q - q' \right\rangle + \tilde{\mathcal{D}}_{\mathrm{KL}}(\sigma(q') \parallel q) &= \langle \sigma(q') - p, q - q' \rangle + \sum_{i\in[d]} \sigma(q')_i \ln \frac{\sigma(q')_i}{\sigma(q)_i} \\ &= \langle \sigma(q') - p, q - q' \rangle + \langle \sigma(q'), q' - q \rangle + \ln \frac{\sum_{j\in[d]} e^{q'_j}}{\sum_{j\in[d]} e^{q_j}}\,,\end{aligned}$$

which completes the proof. □

Proposition A.2. *The Hessian terms in Proposition A.5 is equal to a weighted outer product of gradients:*

$$\frac{\partial^2 \tilde{\mathcal{D}}_{\mathrm{KL}}(q \parallel q)}{\partial q \partial q^\top} = \mathbb{E}_{x\sim\sigma(q)}\left[\left(\frac{\partial \tilde{\mathcal{D}}_{\mathrm{KL}}(\delta_x \parallel q)}{\partial q}\right)\left(\frac{\partial \tilde{\mathcal{D}}_{\mathrm{KL}}(\delta_x \parallel q)}{\partial q}\right)^\top\right]$$

where δ_x is the distribution that is 1 at x and 0 elsewhere.

Proof. By Equation (14), for all $p \in \Delta^d$ and $q \in \mathbb{R}^d$:

$$\frac{\partial^2 \tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel q)}{\partial q \partial q^\top} = \mathrm{diag}(\sigma(q)) - \sigma(q)\sigma(q)^\top .$$

And, by Equation (13), we have

$$\begin{aligned}\mathbb{E}_{x\sim\sigma(q)}\left[\left(\frac{\partial \tilde{\mathcal{D}}_{\mathrm{KL}}(\delta_x \parallel q)}{\partial q}\right)\left(\frac{\partial \tilde{\mathcal{D}}_{\mathrm{KL}}(\delta_x \parallel q)}{\partial q}\right)^\top\right] &= \mathbb{E}_{x\sim\sigma(q)}\left[(\sigma(q)-\delta_x)\,(\sigma(q)-\delta_x)^\top\right] \\ &= \mathrm{diag}(\sigma(q)) - \sigma(q)\sigma(q)^\top .\end{aligned}$$

□

Proposition A.3. *Let Assumption 3.1 holds. Then, we have the following worst-case upper bound for any $\vec{x}$:*

$$\mathcal{D}_{\mathrm{KL}}(p_{\theta^\star}(\cdot \mid \vec{x}) \parallel p_\theta(\cdot \mid \vec{x})) \le 2B + \ln d\,,$$

Proof.

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(p_{\theta^\star}(\cdot \mid \vec{x}) \parallel p_\theta(\cdot \mid \vec{x})) &\le \max_{s\in[d]} -\ln\left(p_\theta(\cdot \mid \vec{x})\right)_s \\ &= -\ln \min_{s\in[d]} \left(p_\theta(\cdot \mid \vec{x})\right)_s \\ &\le -\ln \frac{e^{-B}}{de^{B}} = 2B + \ln d\,,\end{aligned}$$

□

A.2 Self-concordance

We start with the following lemma that provides a useful property of self-concordant functions. This allow us to show that Kullback-Leibler divergence after a softmax link is controlled by a second-order Hessian term around its optimum.

Lemma A.4 ((Bach, 2010, Lemma 1))**.** *Let $r : \mathbb{R} \to \mathbb{R}$ be a convex three times differentiable function such that for all $t \in \mathbb{R}$, $|r'''(t)| \le S r''(t)$, for a fixed $S > 0$. Then, for all $t \ge 0$,*

$$\frac{r''(0)}{S^2}\left(e^{-St} + St - 1\right) \le r(t) - r(0) - tr'(0) \le \frac{r''(0)}{S^2}\left(e^{St} - St - 1\right) .$$

Let us denote $g_{p,q_1,q_2}(t) = \tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel q_1 + tq_2)$ where $p \in \Delta^d$ and $q_1, q_2 \in \mathbb{R}^d$ are arbitrary for now. We omit the subscripts in g whenever it is clear from the context. We compute the first, second, and third order derivatives of the function g:

$$\begin{aligned}g'(t) &= \frac{\partial \tilde{\mathcal{D}}_{\mathrm{KL}}(p \parallel q_1 + tq_2)}{\partial (q_1 + tq_2)} \cdot \frac{\partial (q_1 + tq_2)}{\partial t} = \left\langle (\sigma(q_1 + tq_2) - p)\,, q_2 \right\rangle ,\\ g''(t) &= \left\langle \frac{\partial \sigma(q_1 + tq_2)}{\partial q_1 + tq_2} \cdot \frac{\partial q_1 + tq_2}{\partial t}, q_2 \right\rangle = q_2^\top \Delta\sigma(q_1 + tq_2) q_2\,,\\ g'''(t) &= q_2^\top \left(\frac{\partial \Delta\sigma(q_1 + tq_2)}{\partial t}\right) q_2 = q_2^\top \left(\mathrm{diag}\left(\Delta\sigma(q_1 + tq_2)q_2\right) - 2\Delta\sigma(q_1 + tq_2) q_2 \sigma(q_1 + tq_2)^\top\right) q_2 \\ &= \langle q_2 \odot q_2, \Delta\sigma(q_1 + tq_2)q_2\rangle - 2g''(t)\langle \sigma(q_1 + tq_2), q_2\rangle .\end{aligned}$$

Note that the third-order derivative can be controlled with the second-order derivative as follows:

$$\begin{aligned} |g'''(t)| &\le |\langle q_2 \odot q_2, \Delta\sigma(q_1 + tq_2)q_2\rangle| + 2|g''(t)\langle\sigma(q_1 + tq_2), q_2\rangle| \\ &\le 3\|q_2\|_\infty g''(t)\,. \end{aligned} \tag{15}$$

Proposition A.5. *Let $q, q' \in \mathbb{R}^d$ be such that $\|q\|_\infty \le B, \|q'\|_\infty \le B$ for some constant $B \ge 1$. Set $p = \sigma(q)$. Then, we have the following relation:*

$$c_L(q'-q)^\top \frac{\partial^2 \tilde{\mathcal{D}}_{\mathrm{KL}}(p \,\|\, q)}{\partial q \partial q^\top}(q'-q) \le \tilde{\mathcal{D}}_{\mathrm{KL}}(p \,\|\, q') \le c_U(q'-q)^\top \frac{\partial^2 \tilde{\mathcal{D}}_{\mathrm{KL}}(p \,\|\, q)}{\partial q \partial q^\top}(q'-q)\,,$$

where the derivative is taken only with respect to the second parameter of $\tilde{\mathcal{D}}$ and

$$c_L = \frac{e^{-6B} + 6B - 1}{36B^2} \ge \frac{5}{36B}\,, \quad c_U = \frac{e^{6B} - 6B - 1}{36B^2}\,.$$

Proof. In the construction above, set $p = \sigma(q), q_1 = q, q_2 = q' - q$. Observe that we have $\|q_2\|_\infty \le 2B$. Therefore, by Equation (15), we have:

$$|g'''_{p,q_1,q_2}(t)| \le 6B g''_{p,q_1,q_2}(t)\,.$$

Then, by Lemma A.4, we have

$$c_L g''(0) \le g(1) - g(0) - g'(0) \le c_U g''(0)\,.$$

The result follows by noting that $g(0) = g'(0) = 0, g''(0) = (q'-q)^\top \frac{\partial^2 \tilde{\mathcal{D}}_{\mathrm{KL}}(p \,\|\, q)}{\partial q \partial q^\top}(q'-q)$ and $g(1) = \tilde{\mathcal{D}}_{\mathrm{KL}}(q \,\|\, q')$. □

A.3 The Azuma-Hoeffding Inequality

We use the Azuma-Hoeffding inequality (Azuma, 1967), stated in Theorem A.6, to prove concentration inequalities for empirical processes that appear in our proofs.

Theorem A.6. *(The Azuma-Hoeffding inequality) Let $Y_0, \ldots, Y_n$ be a martingale sequence such that $Y_0 = 0$ and $|Y_i - Y_{i-1}| \le c_i$ for all $i = 1, \ldots, n$. Then, for any $r > 0$,*

$$\mathbb{P}\left(\exists k \ge 0 : Y_k \ge r\right) \le \exp\left(-\frac{r^2}{2\sum_{i=1}^n c_i^2}\right)\,.$$

A.4 Freedman's Inequality

We use the Freedman's inequality, stated in Theorem A.7, to prove concentration inequalities for empirical processes that appear in our proofs. In Lemma A.8, we use a simplification of the result by Yüksel et al. (2024, Lemma B.5.) to compare the quadratic variation with the martingale series itself. This is useful in our proofs to show certain events necessarily implied by empirical risk minimization do not occur with high probability.

Theorem A.7. *(Freedman's inequality) Let $Y_0, \ldots, Y_n$ be a real-valued martingale series that is adapted to the filtration $\mathcal{F}_0, \ldots, \mathcal{F}_n$ where $Y_0 = 0$. Let $d_1, \ldots, d_n$ be the difference sequence induced, i.e.,*

$$d_i = Y_i - Y_{i-1} \quad \text{for } i = 1, \ldots, n.$$

Assume that d_i is upper bounded by some R, i.e., $|d_i| \le R$ for all i. Let W_i be the quadratic variation of the martingale series, i.e.,

$$W_i = \sum_{j=1}^{i} \mathbb{E}[d_j^2 \mid \mathcal{F}_{j-1}] \quad \text{for } i = 1, \ldots, n.$$

Then, for any $r, W > 0$,

$$\mathbb{P}\left(\exists k \ge 0 : Y_k \ge r \ \text{ and } \ W_k \le W\right) \le \exp\left(-\frac{r^2/2}{W + Rr}\right)\,.$$

Lemma A.8. *Let $\gamma > 0, \alpha_U \geq \alpha_L > 0$ be scalars and let $\mathcal{E}$ denote the following event*

$$\mathcal{E} = \{W_n \geq \alpha_L\} \cap \{\gamma Y_n \leq \alpha_U\},$$

where Y_n and W_n verifies the assumptions of Theorem A.7. Then, we have the following concentration inequality

$$\mathbb{P}\left(\{W_n \leq \gamma Y_n\} \cap \mathcal{E}\right) \leq \exp\left(-\frac{\alpha_L}{2e\gamma\left(e\gamma + R\right)} + \ln\left(\ln\left(\frac{\alpha_U}{\alpha_L}\right) + 1\right)\right).$$

Proof. Let $\mathcal{G} = \{\alpha_L, e\alpha_L, \ldots, e^k\alpha_L\}$ where k is the smallest positive integer such that

$$e^k\alpha_L \geq \alpha_U.$$

Then, by a union bound,

$$\begin{aligned}
\mathbb{P}\left(W_n \leq \gamma Y_n \cap \mathcal{E}\right) &\leq \mathbb{P}\left(\cup_{i=1}^k \left(\{W_n \leq e^i\alpha_L, \gamma Y_n \geq e^{i-1}\alpha_L\} \cap \mathcal{E}\right)\right) \\
&\leq \mathbb{P}\left(\cup_{i=1}^k \left(\{W_n \leq e^i\alpha_L, \gamma Y_n \geq e^{i-1}\alpha_L\}\right)\right) \\
&\leq \sum_{i=1}^k \mathbb{P}\left(\{W_n \leq e^i\alpha_L, \gamma Y_n \geq e^{i-1}\alpha_L\}\right).
\end{aligned}$$

By applying Theorem A.7 with $r = e^{i-1}\alpha_L/\gamma$ and $W = e^i\alpha_L$, we obtain

$$\mathbb{P}\left(W_n \leq e^i\alpha_L, Y_n \geq e^{i-1}\alpha_L/\gamma\right) \leq \exp\left(-\frac{e^{i-2}\alpha_L}{2\gamma\left(e\gamma + R\right)}\right),$$

for each $i = 1, \ldots, k$. The union bound gives

$$\sum_{i=1}^k \mathbb{P}\left(\{W_n \leq e^i\alpha_L, Y_n \geq e^{i-1}\alpha_L\}\right) \leq \sum_{i=1}^k \exp\left(-\frac{e^{i-2}\alpha_L}{2\gamma\left(e\gamma + R\right)}\right) \leq \exp\left(-\frac{\alpha_L}{2e\gamma\left(e\gamma + R\right)} + \ln k\right).$$

The result follows by noting that

$$k \leq \ln\left(\frac{\alpha_U}{\alpha_L}\right) + 1.$$

□

Corollary A.9. *Let $\gamma > 0, \alpha_U \geq \alpha_L > 0$ be scalars. Let $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{K}$ denote the following events*

$$\begin{aligned}
\mathcal{E}_1 &= \{\gamma Y_n \leq \alpha_U\}, \quad \mathcal{E}_2 = \{W_n \geq \alpha_L\}, \\
\mathcal{K} &= \{\gamma Y_n \geq \max\{W_n, \alpha_L\}\},
\end{aligned}$$

where Y_n and W_n verifies the assumptions of Theorem A.7. Then, we have the following concentration inequality

$$\mathbb{P}\left(\mathcal{K} \cap \mathcal{E}_1\right) \leq 2\exp\left(-\frac{\alpha_L}{2e\gamma\left(e\gamma + R\right)} + \ln\left(\ln\left(\frac{\alpha_U}{\alpha_L}\right) + 1\right)\right).$$

Proof. By Lemma A.8,

$$\mathbb{P}\left(\mathcal{K} \cap \mathcal{E}_1 \cap \mathcal{E}_2\right) \leq \exp\left(-\frac{\alpha_L}{2e\gamma\left(e\gamma + R\right)} + \ln\left(\ln\left(\frac{\alpha_U}{\alpha_L}\right) + 1\right)\right).$$

By Theorem A.7 with $r = \alpha_L/\gamma$ and $W = \alpha_L$,

$$\mathbb{P}\left(\mathcal{K} \cap \mathcal{E}_2^C\right) \leq \mathbb{P}\left(\gamma Y_n \geq \alpha_L \geq W_n\right) \leq \exp\left(-\frac{\alpha_L}{2\gamma\left(\gamma + R\right)}\right).$$

The result follows by noting that

$$\begin{aligned}
\mathbb{P}\left(\mathcal{K} \cap \mathcal{E}_1\right) &= \mathbb{P}\left(\mathcal{K} \cap \mathcal{E}_1 \cap \mathcal{E}_2\right) + \mathbb{P}\left(\mathcal{K} \cap \mathcal{E}_1 \cap \mathcal{E}_2^C\right) \\
&\leq \mathbb{P}\left(\mathcal{K} \cap \mathcal{E}_1 \cap \mathcal{E}_2\right) + \mathbb{P}\left(\mathcal{K} \cap \mathcal{E}_2^C\right).
\end{aligned}$$

□

B PROOF OF Theorem 4.1

Theorem 4.1. *Let Assumptions 3.1 to 3.3 and 3.5 hold. For any* $0 < \delta < e^{-1}$,

$$\tilde{\mathcal{L}}(\hat{\theta}) = \mathcal{O}\left(\frac{B\mathcal{C}(\Theta)}{NT} \ln eLNT \ln \frac{1}{\delta}\right),$$

with probability at least $1 - \delta$.

Any empirical risk minimizer $\hat{\theta}$ satisfy the following

$$\hat{\mathcal{L}}(\hat{\theta}) \leq \hat{\mathcal{L}}(\theta^\star), \tag{16}$$

by Assumption 3.3, i.e., $\theta^\star \in \Theta$. Equation (16) is the starting point of our analysis. In fact, all of our results applies to any estimators of data, including the empirical risk minimizer, that verifies the condition in Equation (16).

We rewrite the condition in Equation (16) by using Proposition A.1:

$$\hat{Z}(\hat{\theta}) \geq \tilde{\mathcal{L}}(\hat{\theta}), \quad \text{where} \tag{17}$$
$$\hat{Z}(\theta) \coloneqq \frac{1}{NT} \sum_{t,n} \left\langle g_{\theta^\star}(x_t^{(n)}), f_{\theta^\star}(\vec{x}_t^{(n)}) - f_\theta(\vec{x}_t^{(n)}) \right\rangle,$$

where $g_{\theta^\star}(x_t^{(n)})$ denotes the gradient of the ground truth probability with respect to the logits, i.e.,

$$g_{\theta^\star}(x_t^{(n)}) \coloneqq -\partial_{f_{\theta^\star}(\vec{x}_t^{(n)})} \ln p_{\theta^\star}(x_t^{(n)} \mid \vec{x}_t^{(n)}) = p_{\theta^\star}(\vec{x}_t^{(n)}) - \mathcal{E}(x_t^{(n)}).$$

Recall that by Proposition A.5:

$$\frac{5}{36B} \hat{V}(\theta) \leq \tilde{\mathcal{L}}(\theta),$$

where $\hat{V}(\theta)$ is the second-order Taylor approximation of the loss around $\theta^\star$:

$$\begin{aligned}
\hat{V}(\theta) &\coloneqq \frac{1}{NT} \sum_{t,n} \|f_{\theta^\star}(\vec{x}_t^{(n)}) - f_\theta(\vec{x}_t^{(n)})\|^2_{H_{\theta^\star}(\vec{x}_t^{(n)})} \\
&= \frac{1}{NT} \sum_{t,n} \left(f_{\theta^\star}(\vec{x}_t^{(n)}) - f_\theta(\vec{x}_t^{(n)})\right)^\top H_{\theta^\star}(\vec{x}_t^{(n)}) \left(f_{\theta^\star}(\vec{x}_t^{(n)}) - f_\theta(\vec{x}_t^{(n)})\right),
\end{aligned}$$

and $H_{\theta^\star}(\vec{x}_t^{(n)})$ denotes the Hessian of the loss at $\theta^\star$, i.e.,

$$H_{\theta^\star}(\vec{x}_t^{(n)}) \coloneqq \partial_{f_{\theta^\star}(\vec{x}_t^{(n)})} g_{\theta^\star}(\vec{x}_t^{(n)}) = \operatorname{diag}(p_{\theta^\star}(\vec{x}_t^{(n)})) - p_{\theta^\star}(\vec{x}_t^{(n)}) p_{\theta^\star}(\vec{x}_t^{(n)})^\top.$$

Equation (17) implies that the following event is satisfied by $\hat{\theta}$:

$$\{\hat{Z}(\theta) \geq \frac{5}{36B} \hat{V}(\theta)\}, \tag{18}$$

for some constant $\gamma > 0$. Interestingly, Equation (18) does not hold in expectation:

$$\forall \theta \in \Theta, \ \gamma Z(\theta) = 0 < V(\theta), \quad \text{where} \quad Z(\theta) \coloneqq \mathbb{E}\left[\hat{Z}(\theta)\right], \text{ and } V(\theta) \coloneqq \mathbb{E}\left[\hat{V}(\theta)\right].$$

Instead of directly controlling Equation (18), we control the following event:

$$\mathcal{E}_\theta(\hat{\alpha}) \coloneqq \{\hat{Z}(\theta) \geq \frac{5}{36B} \max\{\hat{V}(\theta), \hat{\alpha}\}\}. \tag{19}$$

The proofs of Theorems 4.1, 4.2, 4.4 and 4.6 rely on the following uniform control of Equation (19):

Proposition B.1. *Let $\mathcal{W}(\hat{\alpha})$ denote the following event:*

$$\mathcal{W}(\hat{\alpha}) \coloneqq \{\forall \theta \in \Theta : \mathcal{E}_\theta(\hat{\alpha})\} \,.$$

Then, we have the following lower bound:

$$\mathbb{P}\left(\left\{\tilde{\mathcal{L}}(\hat{\theta}) < \frac{5}{36B}\hat{\alpha}\right\}\right) \geq \mathbb{P}\left(\mathcal{W}(\hat{\alpha})\right) \,.$$

Proof. We show that the right-hand side of the inequality is a subset of the left-hand side. Observe that $\hat{Z}(\theta) < \frac{5}{36B}\max\{\hat{V}(\theta), \hat{\alpha}\}$ and $\tilde{L}(\theta) \leq \hat{Z}(\theta)$ implies that

$$\tilde{L}(\theta) < \frac{5}{36B}\max\{\hat{V}(\theta), \hat{\alpha}\} < \max\{\tilde{L}(\theta), \frac{5}{36B}\hat{\alpha}\}\,,$$

which is possible only if $\tilde{L}(\theta) < \frac{5}{36B}\hat{\alpha}$. □

In order to derive statements on $\mathcal{W}(\hat{\alpha})$, we first control it point-wise. Observe that $NT \cdot \hat{V}(\theta)$ is the *quadratic variation* of $NT \cdot \hat{Z}(\theta)$ by the following relation:

$$\|f_{\theta^\star}(\vec{x}_t^{(n)}) - f_\theta(\vec{x}_t^{(n)})\|^2_{H_{\theta^\star}(\vec{x}_t^{(n)})} = \mathbb{E}_{x_t^{(n)}}\left[\left\langle g_{\theta^\star}(\vec{x}_t^{(n)}), f_{\theta^\star}(\vec{x}_t^{(n)}) - f_\theta(\vec{x}_t^{(n)})\right\rangle^2\right] \,.$$

Note that the martingale errors are bounded for all t, n at every step:

$$\begin{aligned}\left\langle g_{\theta^\star}(\vec{x}_t^{(n)}), f_{\theta^\star}(\vec{x}_t^{(n)}) - f_\theta(\vec{x}_t^{(n)})\right\rangle &= \left\|g_{\theta^\star}(\vec{x}_t^{(n)})\right\|_1 \left\|f_{\theta^\star}(\vec{x}_t^{(n)}) - f_\theta(\vec{x}_t^{(n)})\right\|_\infty \\ &= \left\|p_{\theta^\star}(\vec{x}_t^{(n)}) - \mathcal{E}(\vec{x}_t^{(n)})\right\|_1 \left\|f_{\theta^\star}(\vec{x}_t^{(n)}) - f_\theta(\vec{x}_t^{(n)})\right\|_\infty \leq 4B\,,\end{aligned} \tag{20}$$

where we have used Assumption 3.1. Using Corollary A.9 with $R = 4B, \alpha_L = \hat{\alpha}NT, \alpha_U = 4BNT$ and $\gamma > 0$, we have the following point-wise control:

$$\mathbb{P}\left(\gamma\hat{Z}(\theta) \geq \max\{\hat{V}(\theta), \hat{\alpha}\}\right) \leq 2\exp\left(-\frac{NT\hat{\alpha}}{2e\gamma\left(e\gamma + 4B\right)} + \ln\left(\ln\left(\frac{4B}{\hat{\alpha}}\right) + 1\right)\right) \,, \tag{21}$$

Without loss of generality, assume $\hat{\alpha} \leq 4B$ as otherwise the upper bound in Equation (21) can be replaced with

$$\mathbb{P}\left(\gamma\hat{Z}(\theta) \geq \max\{\hat{V}(\theta), \hat{\alpha}\}\right) = 0\,.$$

To uniformly control $\mathcal{W}(\hat{\alpha})$, we discretize Θ and apply Equation (21) for each element in the discretization. Let $\mathcal{N}_\epsilon(\Theta)$ be an ϵ-net such that for any $\theta \in \Theta$, there exists $\theta' \in \mathcal{N}_\epsilon(\Theta)$ such that $\|\theta - \theta'\|_p \leq \epsilon$. Then, by a union bound, we have the following:

$$\mathbb{P}\left(\left\{\exists\theta \in \mathcal{N}_\epsilon(\Theta) : \gamma\hat{Z}(\theta) \geq \max\{\hat{V}(\theta), \hat{\alpha}\}\right\}\right) \leq 2\exp\left(-\frac{NT\hat{\alpha}}{2e\gamma\left(e\gamma + 4B\right)} + \mathcal{C}_\epsilon(\Theta) + \ln\left(\ln\left(\frac{4B}{\hat{\alpha}}\right) + 1\right)\right) \,.$$

Observe that we have the following implication:

$$\exists\theta \in \Theta : \gamma\hat{Z}(\theta) \geq \max\{\hat{V}(\theta), \hat{\alpha}\} \implies \exists\theta' \in \mathcal{N}_\epsilon(\Theta) : 4\gamma\hat{Z}(\theta') \geq \max\{\hat{V}(\theta'), \hat{\alpha}\}\,.$$

Select $\epsilon > 0$ small enough such that the following holds $\forall\theta \in \mathcal{S}$:

$$\exists\theta' \in \mathcal{N}_\epsilon(\Theta) : \quad \max\{\hat{V}(\theta'), \hat{\alpha}\} \leq \max\{\hat{V}(\theta), \hat{\alpha}\} + \hat{\alpha}\,, \quad \hat{Z}(\theta') \geq \hat{Z}(\theta) - \frac{\hat{\alpha}}{2\gamma}\,. \tag{22}$$

Then, we have the following bound over Θ:

$$\begin{aligned}\mathbb{P}\left(\left\{\forall\theta\in\Theta:\gamma\hat{Z}(\theta)<\max\{\hat{V}(\theta),\hat{\alpha}\}\right\}\right)&=1-\mathbb{P}\left(\left\{\exists\theta\in\Theta:\gamma\hat{Z}(\theta)\geq\max\{\hat{V}(\theta),\hat{\alpha}\}\right\}\right)\\&\geq 1-\mathbb{P}\left(\left\{\exists\theta\in\mathcal{N}_\epsilon(\Theta):4\gamma\hat{Z}(\theta)\geq\max\{\hat{V}(\theta),\hat{\alpha}\}\right\}\right)\\&\geq 1-2\exp\left(-\frac{NT\hat{\alpha}}{32e\gamma\left(e\gamma+B\right)}+\mathcal{C}_\epsilon(\Theta)+\ln\left(\ln\left(\frac{4B}{\hat{\alpha}}\right)+1\right)\right).\end{aligned}\tag{23}$$

Set $\gamma=\dfrac{36B}{5}$ and simplify Equation (23) by $\mathcal{C}_\epsilon(\Theta)\leq\mathcal{C}(\Theta)\ln\dfrac{e}{\epsilon}$ and $x\geq\ln x+1$ for all $x>0$:

$$\mathbb{P}\left(\mathcal{W}\left(\hat{\alpha}\right)\right)\geq 1-2\exp\left(\mathcal{J}(\hat{\alpha};\epsilon)\right),\quad\text{where}\quad\mathcal{J}(\hat{\alpha};\epsilon):=-\frac{25NT\hat{\alpha}}{1152e\left(36e+5\right)B^2}+\mathcal{C}(\Theta)\ln\frac{e}{\epsilon}+\ln\frac{4B}{\hat{\alpha}}.\tag{24}$$

To prove Theorem 4.1, we need to verify that ϵ and $\hat{\alpha}$ can be chosen such that *(i)* ϵ satisfy Equation (22) and *(ii)* $\mathcal{J}(\hat{\alpha};\epsilon)$ is small. Then, the proof of Theorem 4.1 follows from Proposition B.1 and Equation (24).

Choice of ϵ. For the first point, observe that the map

$$\theta\mapsto\left\langle g_{\theta^\star}(\vec{x}_t^{(n)}),f_{\theta^\star}(\vec{x}_t^{(n)})-f_\theta(\vec{x}_t^{(n)})\right\rangle,$$

is Lipschitz continuous with constant $\sqrt{2}L$ in $\|\cdot\|_p$ norm by Assumption 3.2. And, its square is Lipschitz continuous with constant $8\sqrt{2}BL$ where we have used the bound in Equation (20). Then, we have the following Lipschitzness results on $\hat{Z}$ and $\hat{V}$:

$$|\hat{Z}(\theta)-\hat{Z}(\theta')|\leq L\sqrt{2}\|\theta-\theta'\|_p,\quad|\hat{V}(\theta)-\hat{V}(\theta')|\leq 8\sqrt{2}BL\|\theta-\theta'\|_p.$$

Therefore, we can choose ϵ such that

$$\epsilon(\hat{\alpha}):=\min\left\{\frac{5\hat{\alpha}}{72\sqrt{2}BL},\frac{\hat{\alpha}}{8\sqrt{2}BL},1\right\}=\min\left\{\frac{5\hat{\alpha}}{72\sqrt{2}BL},1\right\}.$$

For this particular choice, we denote $\mathcal{J}(\hat{\alpha}):=\mathcal{J}\left(\hat{\alpha},\epsilon\left(\hat{\alpha}\right)\right)$.

Choice of $\hat{\alpha}$. For the latter point on $\mathcal{J}(\hat{\alpha})$, consider the following choice for $\hat{\alpha}$:

$$\hat{\alpha}:=C_{\hat{\alpha}}\frac{B^2\mathcal{C}(\Theta)}{NT}\ln eLNT,\quad\text{where } C_{\hat{\alpha}}=\frac{1152e\left(36e+5\right)}{25}C'_{\hat{\alpha}}\text{ and }C'_{\hat{\alpha}}\text{ is to be chosen later}.\tag{25}$$

Observe that for this choice of $\hat{\alpha}$, we have the following:

$$\epsilon=\min\left\{\frac{5C_{\hat{\alpha}}B\mathcal{C}(\Theta)\ln eLNT}{72\sqrt{2}LNT},1\right\}.$$

Thus, we have

$$\mathcal{J}(\hat{\alpha})=-C'_{\hat{\alpha}}\mathcal{C}(\Theta)\ln eLNT+\mathcal{C}(\Theta)\ln\max\left\{e,\frac{72\sqrt{2}LNT}{5C_{\hat{\alpha}}B\mathcal{C}(\Theta)\ln eLNT}\right\}+\ln\frac{4NT}{C_{\hat{\alpha}}B\mathcal{C}(\Theta)\ln eLNT}.$$

For any confidence level $0<\delta<e^{-1}$, set $C'_{\hat{\alpha}}=2C+\ln\dfrac{1}{\delta}$ where C is a constant that verifies

$$C\mathcal{C}(\Theta)\ln eLNT\geq\mathcal{C}(\Theta)\ln\max\left\{e,\frac{72\sqrt{2}LNT}{5C_{\hat{\alpha}}B\mathcal{C}(\Theta)\ln eLNT}\right\}+\ln\frac{4NT}{C_{\hat{\alpha}}B\mathcal{C}(\Theta)\ln eLNT}.$$

This guarantees $\exp\left(\mathcal{J}(\hat{\alpha})\right)<\delta$ and concludes the proof of Theorem 4.1. Note that the constant $C_{\hat{\alpha}}$ can be chosen such that $C_{\hat{\alpha}}=\mathcal{O}\left(\ln\dfrac{1}{\delta}\right)$.

C PROOF OF Theorem 4.2

Theorem 4.2. *Let Assumptions 3.1 to 3.3 and 3.5 hold. For any* $0 < \delta < e^{-1}$ *and* $T \gg \tau_{\mathrm{mix}} \ln eT$,

$$\mathcal{L}(\hat{\theta}) = \mathcal{O}\left(\tau_{\mathrm{mix}} \frac{(B + \ln d)\, \mathcal{C}(\Theta)}{NT} \ln eLNT \ln eNT \ln \frac{1}{\delta}\right),$$

with probability at least $1 - \delta$.

Star-shaped assumption. In order to prove Theorem 4.2, we first augment the hypothesis class such that $\mathcal{L}$ and $\tilde{\mathcal{L}}$ are star-shaped. This enables fast convergence rates through localization techniques (Wainwright, 2019). More specifically, we set $\Theta' = \Theta \times [0, 1]$ and

$$\forall \theta, \lambda \in \Theta', \forall \vec{x}\,, \quad f_{\theta,\lambda}(\vec{x}) = \lambda'_{\theta,\lambda,\vec{x}} f_\theta(\vec{x}) + (1 - \lambda'_{\theta,\lambda,\vec{x}}) f_{\theta_\star}(\vec{x})\,,$$

where $\lambda'_{\theta,\lambda,\vec{x}} \in [0, 1]$ is chosen such that

$$\tilde{\mathcal{D}}_{\mathrm{KL}}(p_{\theta_\star}(\cdot \mid \vec{x}) \parallel f_{\theta,\lambda}(\vec{x})) = \lambda \tilde{\mathcal{D}}_{\mathrm{KL}}(p_{\theta_\star}(\cdot \mid \vec{x}) \parallel f_\theta(\vec{x}))\,.$$

Such a λ' is guaranteed to exists by the convexity of the Kullback-Leibler divergence. Note that any discretization over Θ can be extended to Θ' by adding the discretization over $[0, 1]$ via Cartesian product. In particular, we consider discretizations over Θ' with a grid of size ϵ over Θ in $\|\cdot\|_p$ and a grid of size ϵ over $[0, 1]$. We extend the definitions of $\hat{Z}, \hat{V}, \tilde{\mathcal{L}}, \mathcal{L}$ to Θ' by replacing f_θ with $f_{\theta,\lambda}$.

Uniform concentration. We follow the argument in Appendix B with an additional step. We prove concentration of $\hat{\mathcal{L}}(\theta, \lambda)$ around $\mathcal{L}(\theta, \lambda)$ uniformly. Let $\mathcal{S}(\alpha; \hat{\alpha})$ denote the following level set

$$\mathcal{S}(\alpha; \hat{\alpha}) \coloneqq \{(\theta, \lambda) \in \Theta' : \mathcal{L}(\theta, \lambda) \geq \alpha + 2\frac{5}{36B}\hat{\alpha}\}\,.$$

In the following, fix $\hat{\alpha}$ to the value set in Equation (25).

We start with the following generalization of Proposition B.1:

Proposition C.1. *Let* $\mathcal{W}(\hat{\alpha})$ *be defined as in Proposition B.1 and let* $\mathcal{B}(\alpha)$ *be the following event:*

$$\mathcal{B}(\alpha) \coloneqq \{\forall (\theta, \lambda) \in \Theta' : |\mathcal{L}(\theta, \lambda) - \tilde{\mathcal{L}}(\theta, \lambda)| \leq \frac{1}{2}\mathcal{L}(\theta, \lambda) + \frac{\alpha}{2}\}\,.$$

Assume that $\alpha > 0$ *verifies the following:*

$$\mathbb{P}(\mathcal{B}(\alpha)) \geq 1 - \delta\,, \tag{26}$$

for some constant $0 < \delta < 1$. *Then, we have the following lower bound:*

$$\mathbb{P}\left(\mathcal{L}(\hat{\theta}) < \alpha + 2\frac{5}{36B}\hat{\alpha}\right) \geq \mathbb{P}(\mathcal{W}(\hat{\alpha})) - \delta\,.$$

Proof. First, note that the following intersection is empty:

$$\mathcal{W}(\hat{\alpha}) \cap \left\{(\hat{\theta}, 0) \in \mathcal{S}(\alpha; \hat{\alpha})\right\} \cap \mathcal{B}(\alpha) = \emptyset\,.$$

Moreover, we have the following by the definition of $\mathcal{S}(\alpha; \hat{\alpha})$ and Proposition A.5:

$$\begin{aligned}
\mathbb{P}\left(\left\{\mathcal{L}(\hat{\theta}) < \alpha + 2\frac{5}{36B}\hat{\alpha}\right\} \cap \left\{(\hat{\theta}, 0) \notin \mathcal{S}(\alpha; \hat{\alpha})\right\} \cap \mathcal{B}(\alpha)\right) &= \mathbb{P}\left(\left\{(\hat{\theta}, 0) \notin \mathcal{S}(\alpha; \hat{\alpha})\right\} \cap \mathcal{B}(\alpha)\right) \\
&\geq \mathbb{P}\left(\mathcal{W}(\hat{\alpha}) \cap \left\{(\hat{\theta}, 0) \notin \mathcal{S}(\alpha; \hat{\alpha})\right\} \cap \mathcal{B}(\alpha)\right) \\
&= \mathbb{P}(\mathcal{W}(\alpha) \cap \mathcal{B}(\alpha)) \\
&\geq \mathbb{P}(\mathcal{W}(\hat{\alpha})) - \mathbb{P}\left(\mathcal{B}(\alpha)^C\right)\,.
\end{aligned}$$

□

The only remaining step for proving Theorem 4.2 is to pick α such that Equation (26) holds. To prove the concentration in Equation (26), we localize the error with Lemma C.2:

Lemma C.2. *(Localization) Let r be a fixed constant. Then, let $\Theta'(r)$ denote the following set:*

$$\Theta'(r) \coloneqq \{(\theta, \lambda) \in \Theta' \mid \mathcal{L}(\theta, \lambda) \leq r\}$$

let $\tilde{S}_r$ denote the following object:

$$\tilde{S}_r = \sup_{(\theta,\lambda)\in\Theta'(r)} \left|\mathcal{L}(\theta, \lambda) - \tilde{\mathcal{L}}(\theta, \lambda)\right| .$$

Then, we have the following relation:

$$\mathbb{P}\left(\mathcal{B}\left(r\right)\right) \geq \mathbb{P}(\tilde{S}_r \leq \frac{r}{2}) \, .$$

Proof. We will show $\mathcal{B}\left(r\right)^C$ is contained in the event $\{\tilde{S}_r > \frac{r}{2}\}$. Assume that there is a $(\theta, \lambda) \in \Theta'$ such that

$$|\mathcal{L}(\theta, \lambda) - \tilde{\mathcal{L}}(\theta, \lambda)| > \frac{1}{2}\mathcal{L}(\theta, \lambda) + \frac{r}{2} \, .$$

If $\mathcal{L}(\theta, \lambda) \leq r$, this is contained in the event $\{\tilde{S}_r > \frac{r}{2}\}$. Otherwise, let $\lambda' = \dfrac{r}{\mathcal{L}(\theta, \lambda)}\lambda$. Then, $\mathcal{L}(\theta, \lambda') = r$ and we have the following for θ':

$$|\mathcal{L}(\theta, \lambda) - \tilde{\mathcal{L}}(\theta, \lambda)| > \frac{1}{2}\mathcal{L}(\theta, \lambda) \implies |\mathcal{L}(\theta, \lambda') - \tilde{\mathcal{L}}(\theta, \lambda')| > \frac{r}{2} \, .$$

□

Independent block method. With Lemma C.2, we only need to prove that

$$\mathbb{P}\left(\tilde{S}_\alpha \leq \frac{\alpha}{2}\right) \geq 1 - \delta \, .$$

If the entries of $\tilde{\mathcal{L}}(\theta)$ are *i.i.d.* random variables, we can use arguments by Wainwright (2019) to control $\tilde{S}_r$. To mimic the *i.i.d.* random variables, we employ the independent block method used by Mohri and Rostamizadeh (2008). Let $\xi \coloneqq \dfrac{\tau_{\text{mix}}\mathcal{C}(\Theta)}{NT}$. Let $Q \coloneqq \tau_{\text{mix}}\left(\xi\right)$ denote the block size and $I \coloneqq \left\lfloor \dfrac{T}{Q} \right\rfloor$ denote the number of blocks. For simplicity, we assume I is divisible by 2 but the proof remains the same if I is odd. Consider the following decomposition:

$$\tilde{\mathcal{L}}(\theta) = \sum_{n=1}^{N}\sum_{i=1}^{I} \tilde{\mathcal{L}}_{i,n}(\theta) \, , \quad \text{where} \quad \tilde{\mathcal{L}}_{i,n}(\theta) \coloneqq \frac{1}{NT} \sum_{j=1}^{\min\{Q, T-(i-1)Q\}} \mathcal{D}_{\text{KL}}(p_{\theta^\star}(\cdot \mid \vec{x}^{(n)}_{(i-1)Q+j}) \parallel p_\theta(\cdot \mid \vec{x}^{(n)}_{(i-1)Q+j})) \, .$$

Then, we have the following upper bound:

$$\tilde{S}_\alpha \leq 2 \sup_{m\in\{1,2\}} \tilde{S}^{(i)}_\alpha \, , \quad \text{where} \quad \tilde{S}^{(m)}_\alpha \coloneqq \sup_{(\theta,\lambda)\in\Theta'(\alpha)} \left| \sum_{n=1}^{N}\sum_{i=0}^{I/2-1} \tilde{\mathcal{L}}_{2i+m,n}(\theta, \lambda) - \frac{\mathcal{L}(\theta, \lambda)}{2} \right| \, ,$$

Hence, it is sufficient to prove

$$\mathbb{P}\left(\tilde{S}^{(i)}_\alpha \leq \frac{\alpha}{4}\right) \geq 1 - \frac{\delta}{2} \, ,$$

for both $i = 1$ and $i = 2$. Since, the proof is the same to both i, we only focus on $\tilde{S}^{(1)}_\alpha$.

Recall that $\vec{x}^{(n)}_1 \sim \pi$ for Theorem 4.2 and π is the stationary distribution of $p_{\theta^\star}$. Thus,

$$\left(x^{(n)}_{-k+1}, \ldots, x^{(n)}_1, \ldots, x^{(n)}_Q\right) \sim \pi_Q \, , \quad \text{where} \quad \pi_Q \coloneqq \mathbb{E}_{\vec{x}_1\sim\pi}\left[p^{(1:k+Q)}_{\theta^\star}\left(\cdot \mid \vec{x}^{(n)}_1\right)\right] \, .$$

Next, we inductively prove that the total variation distance between π_Q and the distribution of $\left(x^{(n)}_{-k+1+2iQ}, \dots, x^{(n)}_{Q+2iQ}\right)$ is small. By the properties of τ_{mix}:

$$\forall \vec{x} \in \mathcal{D}^k\,, \quad \|p^{(Q+1,Q+k)}_{\theta^\star}(\cdot \mid \vec{x}) - \pi\|_{\text{TV}} \le \xi\,.$$

In particular,

$$\left\| p^{(-k+1+2Q,2Q)}_{\theta^\star}\left(\cdot \mid \left(x^{(n)}_{Q-k+1}, \dots, x^{(n)}_{Q}\right)\right) - \pi \right\|_{\text{TV}} \le \xi\,.$$

Therefore,

$$\left(x^{(n)}_{-k+1+2Q}, \dots, x^{(n)}_{3Q}\right) \sim \pi'_Q\,, \quad \text{where} \quad \|\pi'_Q - \pi_Q\|_{\text{TV}} \le \xi\,.$$

Similarly, all $\left(x^{(n)}_{-k+1+2iQ}, \dots, x^{(n)}_{Q+2iQ}\right)$ are ξ-closely distributed to the stationary distribution π_Q.

Let $\bar{\mathcal{L}}_{i,n}(\theta)$ be the *i.i.d.* approximation of $\tilde{\mathcal{L}}_{i,n}(\theta)$:

$$\bar{\mathcal{L}}_{i,n}(\theta) \coloneqq \sum_{j=1}^{\min\{Q,T-(i-1)Q\}} \mathcal{D}_{\text{KL}}(p_{\theta^\star}(\cdot \mid \vec{y}^{(n)}_{(i-1)Q+j}) \parallel p_\theta(\cdot \mid \vec{y}^{(n)}_{(i-1)Q+j}))\,, \quad \text{where} \quad \left(y^{(n)}_{-k+1+(i-1)Q}, \dots, y^{(n)}_{iQ}\right) \sim \pi_Q\,.$$

Let $\bar{S}^{(1)}_\alpha$ be similarly defined to $\tilde{S}^{(1)}_\alpha$ with blocks $\bar{\mathcal{L}}_{i,n}$ instead of $\tilde{\mathcal{L}}_{i,n}$. Then, $\tilde{S}^{(1)}_\alpha$ is controlled by $\bar{S}^{(1)}_\alpha$ as follows:

$$\tilde{S}^{(i)}_\alpha \le \bar{S}^{(i)}_\alpha + \frac{Q\,(2B + \ln d)}{NT} \cdot \frac{N(I-1)}{2}\xi \le \bar{S}^{(i)}_\alpha + \frac{2B + \ln d}{2}\xi\,. \tag{27}$$

Equation (27) holds as each block consist of Q summands bounded in the interval $[0, \frac{2B + \ln d}{NT}]$ by Proposition A.3. As each block of $\tilde{S}^{(1)}_\alpha$ and $\bar{S}^{(1)}_\alpha$ does not agree on only ξ mass, we upper bound the difference with the worst case deviation.

In order to control $\bar{S}^{(1)}_\alpha$, we control the tails of the following quantity defined for $(\theta, \lambda) \in \Theta'(\alpha)$.

$$\bar{S}^{(1)}_\alpha(\theta, \lambda) \coloneqq \sum_{n=1}^{N} \sum_{i=0}^{I/2-1} \bar{\mathcal{L}}_{2i+1,n}(\theta, \lambda) - \frac{\mathcal{L}(\theta, \lambda)}{2}\,.$$

Observe that this is a series with *i.i.d.* entries

$$d_{2i+1,n} \coloneqq \bar{\mathcal{L}}_{2i+1,n}(\theta, \lambda) - \frac{\mathcal{L}(\theta, \lambda)}{NI}\,.$$

$d_{2i+1,n}$ is zero mean due to the stationary, i.e.,

$$\mathbb{E}\left[\bar{\mathcal{L}}_{2i+1,n}(\theta, \lambda)\right] = \mathbb{E}\left[\tilde{\mathcal{L}}_{1,n}(\theta, \lambda)\right] = \frac{1}{NI}\mathbb{E}\left[\tilde{\mathcal{L}}(\theta, \lambda)\right] = \frac{\mathcal{L}(\theta, \lambda)}{NI}\,,$$

and upper bounded in absolute value as follows by Proposition A.3:

$$|d_{2i+1,n}| \le \frac{Q\,(2B + \ln d)}{NT} = \frac{(2B + \ln d)}{NI}\,.$$

Moreover, the variance of $d_{i,n}$ is upper bounded as follows:

$$\mathbb{E}\left[d^2_{2i+1,n}\right] \le \mathbb{E}\left[\bar{\mathcal{L}}_{2i+1,n}(\theta, \lambda)^2\right] \le \frac{Q\,(2B + \ln d)}{NT}\mathbb{E}\left[\bar{\mathcal{L}}_{2i+1,n}(\theta, \lambda)\right] = \frac{Q\,(2B + \ln d)}{NT}\frac{\mathcal{L}(\theta, \lambda)}{NI} \le \frac{(2B + \ln d)\,\alpha}{N^2 I^2}\,.$$

Hence, $W \coloneqq \dfrac{(2B + \ln d)\,r}{2NI}$ upper bounds the quadratic predictable variation of the series. Theorem A.7 then implies

$$\forall \theta \in \Theta(\alpha)\,, \quad \mathbb{P}\left(\bar{S}^{(1)}_\alpha(\theta, \lambda) \ge \frac{\alpha}{8}\right) \le \exp\left(-\frac{\alpha^2/128}{\dfrac{(2B + \ln d)\,\alpha}{4NI} + \dfrac{(2B + \ln d)\,\alpha}{4NI}}\right) \le \exp\left(-\frac{NI\alpha}{64\,(2B + \ln d)}\right)\,.$$

By a uniform bound over the set $\mathcal{N}_\epsilon\left(\Theta'(\alpha)\right)$ with cardinality $|\mathcal{N}_\epsilon\left(\Theta'(\alpha)\right)| \leq \mathcal{C}(\Theta)\ln\frac{e}{\epsilon} + \ln\frac{e}{\epsilon}$, we have

$$\mathbb{P}\left(\exists\theta \in \mathcal{N}_\epsilon\left(\Theta'(\alpha)\right) : \bar{S}_\alpha^{(1)}(\theta) \geq \frac{\alpha}{16}\right) \leq \exp\left(-\frac{NI\alpha}{128\left(2B + \ln d\right)} + 2\mathcal{C}(\Theta)\ln\frac{e}{\epsilon}\right).$$

Using the upper bound in Equation (27) and assuming $\alpha \geq 8\left(2B + \ln d\right)\xi$, we have

$$\mathbb{P}\left(\exists\theta \in \mathcal{N}_\epsilon\left(\Theta'(\alpha)\right) : \tilde{S}_\alpha^{(1)}(\theta) \geq \frac{\alpha}{8}\right) \leq \exp\left(-\frac{NI\alpha}{128\left(2B + \ln d\right)} + 2\mathcal{C}(\Theta)\ln\frac{e}{\epsilon}\right).$$

Recall that $\mathcal{L}(\theta,\lambda)$ and $\tilde{\mathcal{L}}(\theta,\lambda)$ are $\sqrt{2}L$-Lipschitz continuous in $\|\cdot\|_p$ with respect to the first parameter and $1-$*Lipschitz* continuous with respect to the second parameter. We can set ϵ such that

$$\epsilon(\alpha) \coloneqq \frac{1}{32}\min\left\{\frac{\alpha}{\sqrt{2}L}, \alpha\right\},$$

to ensure that

$$\exists(\theta,\lambda) \in \Theta'(\alpha) : \tilde{S}_\alpha^{(1)}(\theta) \geq \frac{\alpha}{4} \implies \exists\theta' \in \mathcal{N}_\epsilon(\Theta'(\alpha)) : \tilde{S}_\alpha^{(1)}(\theta) \geq \frac{\alpha}{8}.$$

Therefore, we have the following:

$$\mathbb{P}\left(\tilde{S}_\alpha^{(1)} \geq \frac{\alpha}{4}\right) \geq \exp\left(-\frac{NI\alpha}{128\left(2B + \ln d\right)} + 2\mathcal{C}(\Theta)\ln\frac{e}{\epsilon}\right). \tag{28}$$

Similar to Appendix B, it is easy to verify that

$$\alpha \gg 128\frac{(2B + \ln d)}{NI}\left(2\mathcal{C}(\Theta)\ln eLNT + \ln\frac{1}{\delta/2}\right),$$

is sufficient to ensure that the right-hand side of Equation (28) is small. The proof follows from Proposition C.1 and Lemma C.2.

D PROOFS OF Theorems 4.4 and 4.6

In order to prove Theorems 4.4 and 4.6, we follow the argument in Appendix B with an additional step. We prove concentration of $\hat{\mathcal{L}}(\theta)$ around $\mathcal{L}(\theta)$ uniformly. Let $\mathcal{S}(\alpha;\hat{\alpha})$ denote the following level set

$$\mathcal{S}(\alpha;\hat{\alpha}) \coloneqq \{\theta \in \Theta : \mathcal{L}(\theta) \geq \alpha + \frac{5}{36B}\hat{\alpha}\}.$$

In the following, we fix $\hat{\alpha}$ to the value set in Equation (25).

We have the following generalization of Proposition B.1:

Proposition D.1. *Let $\mathcal{W}\left(\hat{\alpha}\right)$ be defined as in Proposition B.1 and let $\mathcal{B}\left(\alpha\right)$ be the following event:*

$$\mathcal{B}\left(\alpha\right) = \left\{\forall\theta \in \Theta : \mathcal{L}(\theta) - \tilde{\mathcal{L}}(\theta) \leq \alpha\right\}.$$

Assume that $\alpha > 0$ verifies the following:

$$\mathbb{P}\left(\mathcal{B}\left(\alpha\right)\right) \geq 1 - \delta, \tag{29}$$

for some constant $0 < \delta < 1$. Then, we have the following lower bound:

$$\mathbb{P}\left(\mathcal{L}(\hat{\theta}) < \alpha + \frac{5}{36B}\hat{\alpha}\right) \geq \mathbb{P}\left(\mathcal{W}\left(\hat{\alpha}\right)\right) - \delta.$$

Proof. First, note that the following intersection is empty:

$$\mathcal{W}\left(\hat{\alpha}\right) \cap \left\{\hat{\theta} \in \mathcal{S}\left(\alpha;\hat{\alpha}\right)\right\} \cap \mathcal{B}\left(\alpha\right) = \emptyset.$$

Moreover, we have the following by the definition of $\mathcal{S}(\alpha;\hat{\alpha})$ and Proposition A.5:

$$\begin{aligned}\mathbb{P}\left(\left\{\mathcal{L}(\hat{\theta})<\alpha+\frac{5}{36B}\hat{\alpha}\right\}\cap\left\{\hat{\theta}\notin\mathcal{S}(\alpha;\hat{\alpha})\right\}\cap\mathcal{B}(\alpha)\right)&=\mathbb{P}\left(\left\{\hat{\theta}\notin\mathcal{S}(\alpha;\hat{\alpha})\right\}\cap\mathcal{B}(\alpha)\right)\\&\geq\mathbb{P}\left(\mathcal{W}(\hat{\alpha})\cap\left\{\hat{\theta}\notin\mathcal{S}(\alpha;\hat{\alpha})\right\}\cap\mathcal{B}(\alpha)\right)\\&=\mathbb{P}\left(\mathcal{W}(\alpha)\cap\mathcal{B}(\alpha)\right)\\&\geq\mathbb{P}\left(\mathcal{W}(\hat{\alpha})\right)-\mathbb{P}\left(\mathcal{B}(\alpha)^{C}\right).\end{aligned}$$

□

The only remaining step for proving Theorems 4.4 and 4.6 is to pick α such that Equation (29) holds. To do so, we prove the following discretized version:

$$\mathbb{P}\left(\mathcal{B}_{\epsilon}(\alpha)\right)\geq 1-\delta\,,\quad\text{where}\quad\mathcal{B}_{\epsilon}(\alpha)\coloneqq\left\{\forall\theta\in\mathcal{N}_{\epsilon}(\Theta):\mathcal{L}(\theta)-\tilde{\mathcal{L}}(\theta)\leq\frac{\alpha}{2}\right\},\tag{30}$$

for some arbitrary constant $0<\delta<e^{-1}$. Recall that $\mathcal{L}$ and $\tilde{\mathcal{L}}$ are $\sqrt{2}L$-Lipschitz continuous in $\|\cdot\|_p$ norm by Assumption 3.2. We can set ϵ such that

$$\epsilon(\alpha)\coloneqq\min\left\{\frac{\alpha}{4\sqrt{2}L},1\right\},\tag{31}$$

to ensure that

$$\exists\theta\in\Theta:\mathcal{L}(\theta)-\tilde{\mathcal{L}}(\theta)>\alpha\implies\exists\theta'\in\mathcal{N}_{\epsilon}(\Theta):\mathcal{L}(\theta')-\tilde{\mathcal{L}}(\theta')>\frac{\alpha}{2}\,.$$

Therefore, the lower bound in Equation (30) implies Equation (29) with the same constant δ.

D.1 Proof of Theorem 4.4

Theorem 4.4. *Let Assumptions 3.1 to 3.3 and 3.5 hold. For any $0<\delta<e^{-1}$,*

$$\begin{aligned}\mathcal{L}(\hat{\theta})\leq\mathcal{O}\Bigg(&\frac{B\mathcal{C}(\Theta)}{NT}\ln eLNT\ln\frac{1}{\delta}+\\&\tilde{\tau}_{\text{mix}}\left(\frac{1}{T}\right)(B+\ln d)\sqrt{\frac{\mathcal{C}(\Theta)}{NT}\ln eLNT\ln\frac{1}{\delta}}\Bigg),\end{aligned}$$

with probability at least $1-\delta$.

Let α be as follows:

$$\alpha\coloneqq 2\sqrt{2}C_{\alpha}\tilde{\tau}_{\text{mix}}\left(\frac{1}{T}\right)(2B+\ln d)\sqrt{\frac{\mathcal{C}(\Theta)}{NT}\ln eLNT}\,,$$

where $C_{\alpha}\geq 1$ is chosen the same as in Appendix D.2.

The proof is same as in Appendix D.2, with the following difference: the upper bound $k\eta(T)$ in Equation (36) is replaced by $\tilde{\tau}_{\text{mix}}\left(\frac{1}{T}\right)$. Once this is proven, Equation (30) holds for a $C_{\alpha}=\mathcal{O}\left(\sqrt{\ln\frac{1}{\delta}}\right)$ with the same argument where $\tilde{\tau}_{\text{mix}}\left(\frac{1}{T}\right)$ replaces $k\eta(T)$.

We only need to prove that for any $(\vec{x},\vec{y})\in\mathcal{X}$, there exist a coupling γ such that

$$\mathbb{E}_{\vec{x}_{(k+1:T+k)},\vec{y}_{(k+1:T+k)}\sim\gamma}\left[\sum_{i=k+1}^{T+k+1}\mathbb{1}\left(\vec{x}_{(i-k:i-1)}\neq\vec{y}_{(i-k:i-1)}\right)\right]\leq\tilde{\tau}_{\text{mix}}\left(\frac{1}{T}\right)+1\,.$$

Assume that $t_{\text{mix}}\coloneqq\tilde{\tau}_{\text{mix}}\left(\frac{1}{T}\right)<T$ as otherwise, the result is trivial.

By the definition of $\tilde{\tau}_{\text{mix}}$, we have the following:

$$\|p_{\theta^\star}^{(t_{\text{mix}}+1,t_{\text{mix}}+k)}(\cdot \mid \vec{x}) - p_{\theta^\star}^{(t_{\text{mix}}+1,t_{\text{mix}}+k)}(\cdot \mid \vec{y})\|_{\text{TV}} \le \frac{1}{T}\,. \tag{32}$$

Note that Equation (32) also upper bounds the TV distance between the marginals of $\vec{x}_{(t_{\text{mix}}+1:T+k)}$ and $\vec{y}_{(t_{\text{mix}}+1:T+k)}$. Let γ_1 be any maximal coupling between $\vec{x}_{(t_{\text{mix}}+1:T+k)}$ and $\vec{y}_{(t_{\text{mix}}+1:T+k)}$, verifying

$$\mathbb{P}_{\vec{x}_{(t_{\text{mix}}+1:T+k)},\vec{y}_{(t_{\text{mix}}+1:T+k)}\sim\gamma_1}\left(\vec{x}_{(t_{\text{mix}}+1:T+k)} \neq \vec{y}_{(t_{\text{mix}}+1:T+k)}\right) = \|p_{\theta^\star}^{(t_{\text{mix}}+1:T+k)}(\cdot \mid \vec{x}) - p_{\theta^\star}^{(t_{\text{mix}}+1:T+k)}(\cdot \mid \vec{y})\|_{\text{TV}}\,.$$

Let γ_2 be the conditional product coupling between $\vec{x}_{(k+1:t_{\text{mix}})}$ and $\vec{y}_{(k+1:t_{\text{mix}})}$, verifying

$$\gamma_2\left(\vec{x}_{(k+1:t_{\text{mix}})},\vec{y}_{(k+1:t_{\text{mix}})};\vec{x}_{(t_{\text{mix}}+1:T+k)},\vec{y}_{(t_{\text{mix}}+1:T+k)}\right) \coloneqq \mathbb{P}\left(\vec{x}_{(k+1:t_{\text{mix}})} \mid \vec{x}_{(t_{\text{mix}}+1:T+k)}\right)\mathbb{P}\left(\vec{y}_{(k+1:t_{\text{mix}})} \mid \vec{y}_{(t_{\text{mix}}+1:T+k)}\right)\,.$$

Then, let γ be the following coupling:

$$\gamma\left(\vec{x}_{(k+1:T+k)},\vec{y}_{(k+1:T+k)}\right) = \gamma_1\left(\vec{x}_{(t_{\text{mix}}+1:T+k)},\vec{y}_{(t_{\text{mix}}+1:T+k)}\right)\gamma_2\left(\vec{x}_{(k+1:t_{\text{mix}})},\vec{y}_{(k+1:t_{\text{mix}})};\vec{x}_{(t_{\text{mix}}+1:T+k)},\vec{y}_{(t_{\text{mix}}+1:T+k)}\right)\,.$$

For this choice of γ, we have the desired upper bound:

$$\begin{aligned}
&\mathbb{E}_{\vec{x}_{(k+1:T+k)},\vec{y}_{(k+1:T+k)}\sim\gamma}\left[\sum_{i=k+1}^{T+k+1}\mathbb{1}\left(\vec{x}_{(i-k:i-1)} \neq \vec{y}_{(i-k:i-1)}\right)\right] \\
&\qquad\le \mathbb{E}_{\vec{x}_{(t_{\text{mix}}+1:T+k)},\vec{y}_{(t_{\text{mix}}+1:T+k)}\sim\gamma_1}\left[\sum_{i=t_{\text{mix}}+k+1}^{T+k+1}\mathbb{1}\left(\vec{x}_{(i-k:i-1)} \neq \vec{y}_{(i-k:i-1)}\right)\right] + (t_{\text{mix}}-k) \\
&\qquad\le \frac{1}{T}\left(T - t_{\text{mix}} + 1\right) + (t_{\text{mix}} - k) \le t_{\text{mix}}\,.
\end{aligned}$$

D.2 Proof of Theorem 4.6

Theorem 4.6. *Let Assumptions 3.1 to 3.3, 3.5 and 4.5 hold. For any $0 < \delta < e^{-1}$,*

$$\begin{aligned}
\mathcal{L}(\hat{\theta}) \le \mathcal{O}\Bigg(&\frac{B\mathcal{C}(\Theta)}{NT}\ln eLNT\ln\frac{1}{\delta} + \\
&k\eta(T)\left(B + \ln d\right)\sqrt{\frac{\mathcal{C}(\Theta)}{NT}\ln eLNT\ln\frac{1}{\delta}}\Bigg)\,,
\end{aligned}$$

with probability at least $1-\delta$.

Let α be denoted as follows:

$$\alpha \coloneqq 2\sqrt{2}C_\alpha k\eta(T)\left(2B + \ln d\right)\sqrt{\frac{\mathcal{C}(\Theta)}{NT}\ln eLNT}\,, \tag{33}$$

where $C_\alpha \ge 1$ to be chosen later. Recall that we only need to prove Equation (30) for a $C_\alpha = \mathcal{O}(\ln\frac{1}{\delta})$.

Consider the following ordering of tokens:

$$x_1^{(1)},\ldots,x_T^{(1)},x_1^{(2)},\ldots,x_T^{(2)},\ldots,x_1^{(N)},\ldots,x_T^{(N)}\,. \tag{34}$$

Let $\mathcal{F}_i$ be the sigma-algebra created by the first i tokens of the sequence in Equation (34). Then, the deviation of $\tilde{L}(\theta)$ from its expectation $L(\theta)$ can be controlled with the following decomposition:

$$L(\theta) - \tilde{L}(\theta) = \sum_{i=1}^{NT} d_i\,, \quad \text{where} \quad d_i \coloneqq \mathbb{E}\left[\tilde{L}(\theta) \mid \mathcal{F}_{i-1}\right] - \mathbb{E}\left[\tilde{L}(\theta) \mid \mathcal{F}_i\right]\,. \tag{35}$$

Note that d_i is zero mean and the series $\{d_i\}_{i=1}^{NT}$ is a martingale difference series.

We upper bound the martingale differences d_i in Equation (35) by using the local mixing time $\tilde{\tau}_{\text{mix}}$ as follows. Let $m \in [NT]$ be an index. Let $n \coloneqq \lceil \frac{m}{T} \rceil$ be the sequence index of $m \in [NT]$ and let $j \coloneqq i \bmod T$ denote its position on its sequence. Then, let $\vec{x}$ denote the following context:

$$\vec{x} \coloneqq \begin{cases} \left(x_{j-k+1}^{(n)}, \dots, x_j^{(n)}\right), & j \geq k\,, \\ \left(\phi, \dots, \phi, x_1^{(n)}, \dots x_j^{(n)}\right), & j < k\,. \end{cases}$$

Consider any $\vec{y}$ such that $(\vec{x}, \vec{y}) \in \mathcal{X}$. Then, by Assumption 4.5, there exist a coupling γ of $\vec{x}_{(k+1:T+k)}$ and $\vec{y}_{(k+1:T+k)}$ such that

$$\mathbb{E}_{\vec{x}_{(k+1:T+k)}, \vec{y}_{(k+1:T+k)} \sim \gamma} \left[\sum_{i=k+1}^{T+k+1} \mathbb{1}\left(\vec{x}_{(i-k:i-1)} \neq \vec{y}_{(i-k:i-1)}\right) \right] \leq k\eta(T)\,. \tag{36}$$

The additional k term pops up due to the fact that each difference in Assumption 4.5 leads to at most k contexts with differences.

Observe that the two terms $\mathbb{E}\left[\tilde{L}(\theta) \mid \mathcal{F}_m\right]$ and $\mathbb{E}\left[\tilde{L}(\theta) \mid \mathcal{F}_{m-1}\right]$ in d_m share the same expectations for sequences $n' \in [N] \neq n$. In addition, the summands with entries from the initial tokens, $t \in [T] < j-1$, cancel out. These two observations lead to the following upper bound:

$$\begin{aligned} d_m \leq \frac{1}{NT} \sup_{\vec{y} \text{ s.t. } (\vec{x},\vec{y}) \in \mathcal{X}} \mathbb{E}_{\vec{x}_{(k+1:T-j+k)}, \vec{y}_{(k+1:T-j+k)} \sim \gamma} \Bigg[\sum_{i=k+1}^{T-j+k+1} & \mathcal{D}_{\text{KL}}(p_{\theta^\star}(\cdot \mid \vec{x}_{(i-k:i-1)}) \parallel p_\theta(\cdot \mid \vec{x}_{(i-k:i-1)})) \\ & - \mathcal{D}_{\text{KL}}(p_{\theta^\star}(\cdot \mid \vec{y}_{(i-k:i-1)}) \parallel p_\theta(\cdot \mid \vec{y}_{(i-k:i-1)})) \Bigg]\,, \end{aligned} \tag{37}$$

where we use the fact that

$$d_m = \mathbb{E}\left[\tilde{\mathcal{L}}(\theta) \mid \mathcal{F}_m\right] - \mathbb{E}\left[\tilde{\mathcal{L}}(\theta) \mid \mathcal{F}_{m-1}\right] \leq \sup_{m'} \left(\mathbb{E}\left[\tilde{\mathcal{L}}(\theta) \mid \mathcal{F}_m\right] - \mathbb{E}\left[\tilde{\mathcal{L}}(\theta) \mid \mathcal{F}_{m'}\right] \right)\,,$$

with $\{\mathcal{F}_{m'}\}$ denoting the set of sigma-algebras where the last token of $\mathcal{F}_m$ is replaced by any token from $\mathcal{D}$. Equation (36) guarantees there are at most $k\eta(T)$ indices in Equation (37) which $\vec{x}_{(i-k:i-1)}$ and $\vec{y}_{(i-k:i-1)}$ differ.

Finally, by bounding every summand with its absolute value, we obtain

$$\forall m \in [NT]\,, \quad d_m \leq \frac{k\eta(T)\,(2B + \ln d)}{NT}\,,$$

where we use Proposition A.3 to upper bound the summands in Equation (37).

The martingale series in Equation (35) then can be controlled by Theorem A.6. The quadratic differences are upper bounded as follows:

$$\sum_{i=1}^{NT} |d_i|^2 \leq \sum_{i=1}^{NT} \left(\frac{k\eta(T)\,(2B + \ln d)}{NT} \right)^2 \leq \frac{(k\eta(T))^2\,(2B + \ln d)^2}{NT}\,,$$

by the bound on each $|d_i|$ above. Therefore, by Theorem A.6 with $r = \sqrt{2} C_\alpha k\eta(T)\,(2B + \ln d) \sqrt{\frac{\mathcal{C}(\Theta)}{NT} \ln eLNT}$,

$$\mathbb{P}\left(\mathcal{L}(\theta) - \tilde{\mathcal{L}}(\theta) \geq \sqrt{2} C_\alpha k\eta(T)\,(2B + \ln d) \sqrt{\frac{\mathcal{C}(\Theta)}{NT} \ln eLNT} \right) \leq \exp\left(-C_\alpha^2 \mathcal{C}(\Theta) \ln eLNT\right)\,.$$

By a uniform bound,

$$\begin{aligned} \mathbb{P}\Bigg(\exists \theta \in \mathcal{N}_\epsilon(\Theta)\,, \mathcal{L}(\theta) - \tilde{\mathcal{L}}(\theta) \geq \sqrt{2} C_\alpha k\eta(T)\,(2B + \ln d) \sqrt{\frac{\mathcal{C}(\Theta)}{NT} \ln eLNT} \Bigg) \\ \leq \exp\left(-C_\alpha^2 \mathcal{C}(\Theta) \ln eLNT + \mathcal{C}(\Theta) \ln \frac{e}{\epsilon} \right)\,. \end{aligned}$$

Similar to Appendix B, the choice of ϵ in Equation (31) ensures that the right-hand side is upper bounded by δ whenever $C_\alpha^2 \gg \ln \frac{1}{\delta}$. Therefore, there exist a $C_\alpha = \mathcal{O}\left(\sqrt{\ln \frac{1}{\delta}}\right)$ that $\mathcal{S}$ defined in Equation (33) satisfies Equation (30).