# Generalization Bounds for Autoregressive Processes and In-Context Learning

**Oğuz Kaan Yüksel**[*], **Nicolas Flammarion**
Theory of Machine Learning Lab,
EPFL, Switzerland

## Abstract

In this paper, we derive generalization results for next-token risk minimization in autoregressive processes of unbounded order. Our starting point is to relate the empirical loss to the denoising loss, which requires no additional assumptions compared to fixed-order Markovian models. We then show that, under a mixing or rephrasability condition on the data-generating process and assuming a stable hypothesis class, the out-of-sample generalization error concentrates around the denoising error. These results characterize sample complexity in terms of the number of tokens, rather than the number of i.i.d. sequences. As a primary application, we interpret in-context learning as a special case of autoregressive prediction and derive sample complexity bounds under similar conditions. Importantly, the properties of individual in-context tasks determine the generalization rates, without requiring assumptions on mixture processes. This perspective suggests that in-context learning can exploit the task decomposition to learn efficiently.

## 1 Introduction

In-context learning—adapting to new tasks using a few input-output examples in the prompt without updating model parameters—has attracted significant attention [Olsson et al., 2022]. Recent work views models as performing inference over the prompt and focuses on interpreting these implicit inference algorithms. Xie et al. [2022] posits that in-context learning is an implicit form of Bayesian inference. While these perspective clarifies capabilities and limitations, it sheds little light on generalization properties.

Statistical learning theory has traditionally focused on i.i.d. data [Vapnik and Chervonenkis, 2015]. Extensions to dependent data have mainly assumed stationary and mixing sequences [Yu, 1994], leading to generalization bounds for non-i.i.d. settings [Mohri and Rostamizadeh, 2008, 2010]. However, these frameworks cannot accommodate autoregressive processes whose order grows with sequence length. The recent work of Li et al. [2023] provides a first step in this direction for in-context learning via stability [Bousquet and Elisseeff, 2002], but it is limited to in-context tasks with i.i.d. data or simple first-order dynamical systems.

We address these gaps by developing a generalization framework for autoregressive models, also accommodating in-context learning. Our theory allows for unbounded dependencies under mixing or rephrasability condition on the data-generating process. Remarkably, for in-context learning, such conditions are only needed at the task level, with no assumptions on the training-induced mixture process. This reveals that autoregressive processes decomposable into simple tasks can be learned efficiently with in-context learning, suggesting that it can be seen as an efficient learning strategy for complex autoregressive tasks.

[*] Correspondence to `oguz.yuksel@epfl.ch`.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Principles of Generative Modeling (PriGM) @ EurIPS2025.

## 2 Problem Setting

**Function class.** Let $\mathcal{D} = t_1, \ldots, t_d$ be a dictionary of discrete input tokens, where $d \in \mathbb{N}^*$ denotes the number of distinct tokens. We consider learning autoregressive processes over $\mathcal{D}^\star := \cup_{i=1}^{\infty} \mathcal{D}^i$ and treat elements of $\mathcal{D}^\star$ as vectors $\vec{x} = (x_1, \ldots, x_k) \in \mathcal{D}^k$. Let $\mathcal{F} := \{f_\theta : \mathcal{D}^\star \to \mathbb{R}^d \,|\, \theta \in \Theta\}$ be a class that maps sequences from $\mathcal{D}^\star$ to an output logit space in $\mathbb{R}^d$. For each $f_\theta$ and sequence $\vec{x} \in \mathcal{D}^\star$, let $p_\theta(\cdot|\vec{x}) = \sigma\left(f_\theta\left(\vec{x}\right)\right) \in \mathcal{P}(\mathcal{D})$ denote the probability distribution that is induced by $f_\theta$, where $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ is the softmax activation function. $\mathcal{F}$ then defines a class of autoregressive models: $\mathcal{P} := \{p_\theta : \mathcal{D}^\star \to \mathcal{P}(\mathcal{D}) \,|\, \theta \in \Theta\}$.

The classical example of such a class is ergodic Markov chains of order $p$ where the predictions at time $t$ depend only on the previous $p$ tokens. Transformers [Vaswani et al., 2017] or state space models [Gu et al., 2021] are modern examples with full context dependency.

**Empirical risk minimization.** We assume the following data generation procedure for the sequences. The sequences are initialized by sampling a prompt $\vec{z} = (z_1, \ldots, z_P)$ from a distribution $\pi^\star$ over $\mathcal{D}^P$ where $P \in \mathbb{N}$ is fixed. Then, full sequences are generated autoregressively by the ground truth $p^\star$: $x_t \sim p^\star(\cdot \mid \vec{x}_t)$, where $\vec{x}_t := (z_1, \ldots, z_P, x_1, \ldots, x_{t-1})$. We assume access to $N$ training sequences of length $T$ generated i.i.d. from $\pi^\star$ and $p^\star$:

$$\forall n \in [N] : \quad \vec{z}^{(n)} := \left(z_1^{(n)}, \ldots, z_P^{(n)}\right) \overset{\text{i.i.d}}{\sim} \pi^\star , \quad \forall n \in [N], t \in [T] : \quad x_t^{(n)} \sim p^\star(\cdot|\vec{x}_t^{(n)}) ,$$

where $\vec{x}_t^{(n)} := (z_1^{(n)}, \ldots, z_P^{(n)}, x_1^{(n)}, \ldots, x_{t-1}^{(n)})$. To learn an estimate $p_{\hat{\theta}}$ of the ground truth $p^\star$, we minimize the negative log likelihood $\mathcal{L}_{\text{train}}(\theta) := \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_{\text{train}}^{(n)}(\theta)$ with $\mathcal{L}_{\text{train}}^{(n)}(\theta) := \frac{1}{T}\sum_{t=1}^{T} \ell_\theta(\vec{x}_t^{(n)}; x_t^{(n)})$ and $\ell_\theta(\vec{x}; x) := -\log p_\theta(x \mid \vec{x}) - \mathcal{H}\left(p^\star(\cdot \mid \vec{x})\right)$ is the shifted cross entropy loss. Given a minimizer $\hat{\theta} \in \Theta$ that verifies $\mathcal{L}_{\text{train}}(\hat{\theta}) - \inf_{\theta \in \Theta} \mathcal{L}_{\text{train}}(\theta) \leq \epsilon_{\text{opt}}$ for some $\epsilon_{\text{opt}}$, we are interested in the generalization error of the model $\hat{\theta}$.

**Generalization errors.** We are interested in in-sample or out-of-sample prediction error:

$$\mathcal{L}_{\text{in}}(\theta) := \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_{\text{in}}^{(n)}(\theta) := \frac{1}{N}\sum_{n=1}^{N} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t^{(n)}}\left[\ell_\theta(\vec{x}_t^{(n)}; x_t^{(n)})\right] , \quad \mathcal{L}_{\text{out}}(\theta) := \mathbb{E}_{\vec{x}_T^{(1)}}\left[\mathcal{L}_{\text{in}}^{(1)}(\theta)\right] .$$

In-sample prediction error studied by Foster et al. [2020], Lotfi et al. [2024] measures the denoising error over the training trajectories, while out-of-sample prediction error measures the generalization error over all trajectories. If $p^\star$ has no specific structure, these two generalization errors can differ significantly as the dynamics over the training trajectories is not always informative about the dynamics of the test trajectories [Tsiamis et al., 2023]. Therefore, we expect the in-sample prediction error $\mathcal{L}_{\text{in}}$ to be small regardless of the properties of $p^\star$ but $\mathcal{L}_{\text{out}}$ to be small only when the training set is representative of the test set.

## 3 Learning Autoregressive Processes

We assume all functions $f \in \mathcal{F}$ satisfy the boundedness condition:

**Assumption 3.1** (Bounded logits). *For all $\theta \in \Theta$, $\sup_{\vec{z} \in \text{supp}(\pi^\star)} \sup_{\vec{x} \in \mathcal{D}^\star} \|f_\theta(\vec{z} \circ \vec{x})\|_\infty \leq B$.*

Proposition E.7 ensures that the loss at each step is controlled by $B$, which is standard in the literature [Mohri and Rostamizadeh, 2008]. Furthermore, to derive fast rates, we use how well the function class $\mathcal{P}$ uniformly approximates $p^\star$:

**Definition 3.2** (Near well-specification). *Let $\epsilon_{\text{app}}$ be the the following constant*

$$\epsilon_{\text{app}} := \min_{\theta} \sup_{\vec{z} \in \text{supp}(\pi^\star)} \sup_{\vec{x} \in \mathcal{D}^\star} \|p^\star(\cdot|\vec{z} \circ \vec{x}) - p_\theta(\cdot|\vec{z} \circ \vec{x})\|_{\text{TV}} .$$

Our results depend on the sequential metric entropy [Rakhlin and Sridharan, 2015, Bilodeau et al., 2020, Jia et al., 2025] induced by the norm $\|f\|_{2,\infty} := \sup_{u \in \text{dom}(f)} \|f(u)\|_2$:

**Definition 3.3** (Sequential metric entropy). *The sequential metric entropy is defined as:*

$$\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) := \mathcal{H}_{2,\infty}(\mathcal{F}|_{U_T}, \epsilon) , \quad U_T = \cup_{i=1}^{T} \mathcal{D}^i \text{ and } \mathcal{F}|_U := \{f|_U \mid f \in \mathcal{F}\} , \qquad (1)$$

*where $\mathcal{H}_p(\mathcal{F}, \epsilon)$ is the metric entropy of $\mathcal{F}$ with norm $p$. For brevity, set $\mathcal{C}(\epsilon) := \mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T)$.*

Lastly, we state our bounds in terms of $\epsilon$ that implicitly verifies a condition regarding the metric entropy. We discuss the metric entropy and $\epsilon$ for various function classes in Section I.

### 3.1 In-sample generalization

We first prove the in-sample generalization for the autoregressive setting:

**Theorem 3.4** (In-sample generalization)**.** *Let Assumption 3.1 hold. With probability* $1-\delta$*,*

$$\mathcal{L}_{\mathrm{in}}(\hat{\theta})-\min_{\theta\in\Theta}\mathcal{L}_{\mathrm{in}}(\theta)=\tilde{\mathcal{O}}\left(\min\left\{\sqrt{\frac{\tilde{B}^2\mathcal{C}(\epsilon)+\log\frac{e}{\delta}}{NT}}+\epsilon_{\mathrm{opt}},\max\left\{\frac{B\mathcal{C}(\epsilon)+\log\frac{e}{\delta}}{NT},B\epsilon_{\mathrm{app}}+\epsilon_{\mathrm{opt}}\right\}\right\}\right).$$

Theorem 3.4 shows that as long as $\epsilon_{\mathrm{opt}}$ is small, the generalization error is bounded by the ratio of complexity of the class over the number of tokens. Notably, if Definition 3.2 holds with small constant, the rate matches the fast decay rate $1/NT$ up to logarithmic factors, matching the i.i.d. discrete distribution setting. These decay rates do not rely on assumptions on the data-generating process $p^\star$. Therefore, it applies to learning from any class that verifies the mild condition in Assumption 3.1. In particular, it implies that over the training trajectories, the temporal dependencies are not detrimental to learning, which has crucial implications for memorization [Carlini et al., 2023].

### 3.2 Out-of-sample generalization

We start with the simplest i.i.d. rate based on the number of independent sequences:

**Theorem 3.5** (i.i.d. rate)**.** *Let Assumption 3.1 hold. With probability* $1-\delta$*,*

$$\sup_{\theta\in\Theta}\left|\mathcal{L}_{\mathrm{out}}(\theta)-\mathcal{L}_{\mathrm{in}}(\theta)\right|=\tilde{\mathcal{O}}\left(\min\left\{\sqrt{\frac{\tilde{B}^2\mathcal{C}(\epsilon)+\log\frac{e}{\delta}}{N}},\frac{\tilde{B}\mathcal{C}(\epsilon)+\log\frac{e}{\delta}}{N}+B\epsilon_{\mathrm{app}}\right\}\right).$$

Theorem 3.5 does not yield an upper bound that accounts for $T$. In the context of pretraining of language models, this is a huge inefficiency as $T$ can be very large, e.g., a book or a long piece of code. To make use of number of tokens in each sequence, we rely on the stability of the function class and the mixing property of the data generation process.

**Definition 3.6** ($K$-stability)**.** *Let* $\mathcal{X}:=\{(\vec{x},\vec{y})\,|\,\vec{x}\in\mathcal{D}^T,|\vec{x}|=|\vec{y}|,\exists!i\in[|\vec{x}|]:x_i\neq y_i\}$*. We say that* $\theta\in\Theta$ *is* $K$*-stable if* $\sup_{\vec{z}\in\mathrm{supp}(\pi^\star)}\sup_{\vec{x},\vec{y}\in\tilde{\mathcal{X}}}\sum_{i=1}^T\|f_\theta(\vec{z}\circ\vec{x}_{1:i})-f_\theta(\vec{z}\circ\vec{y}_{1:i})\|_2\le K$*.*

Assumption 3.7 relates to the stability assumption in [Li et al., 2023], which itself connects to the classical stability notion from [Bousquet and Elisseeff, 2002]. We require the estimator and the minimizer to be in the $K$-stable set with high probability:

**Assumption 3.7.** *Let* $\Theta_K:=\{\theta\in\Theta:\theta\text{ is }K-\text{stable}\}$*. There exist a* $\delta_1>0$ *such that*

$$\mathbb{P}\left(\{\hat{\theta}\in\Theta_K\}\right)\ge 1-\delta_1\quad\textit{and}\quad\mathrm{argmin}_\theta\,\mathcal{L}(\theta)\subset\Theta_K.$$

Given a prompt $\vec{z}$, let $\mathcal{P}_{\vec{z}}^{k_1:k_2}$ denote the law of the $k_1$-th to $k_2$-th token following $\vec{z}$, respectively. Similarly, let $\mathcal{P}^{k_1:k_2}$ denote the average laws induced by sampling of $\vec{z}\sim\pi^\star$. Finally, let $\mathcal{F}_{\vec{z}}^k$ and $\mathcal{F}^k$ be $\sigma$-algebras of $\mathcal{P}_{\vec{z}}^{1:k}$ and $\mathcal{P}^{1:k}$, respectively.

**Definition 3.8** ($\phi$-mixing)**.** *Let* $\phi(\epsilon)$ *be the smallest* $s\in\mathbb{N}$ *s.t.* $\sup_{\vec{z}\in\mathrm{supp}(\pi^\star)}\phi_{\vec{z}}(s)\le\epsilon$ *where*

$$\phi_{\vec{z}}(s):=\sup_{k\in\mathbb{N}}\sup_{A\in\mathcal{F}_{\vec{z}}^k}\|\mathcal{P}_{\vec{z}}^{k+s:\infty}(\cdot\mid A)-\mathcal{P}^{k+s:\infty}(\cdot)\|_{\mathrm{TV}}.$$

Definition 3.8 measures the memory of the process. A fast-decaying $\phi(\epsilon)$ implies that the process forgets its past quickly and we expect learning to happen closer to the i.i.d. setting. A classical example of such processes is ergodic Markov chains with a fixed order. We obtain the following out-of-sample generalization error:

**Theorem 3.9** (Out-of-sample generalization)**.** *Let Assumptions 3.1 and 3.7 hold. Let* $\rho := \min_{\epsilon \in [0,1]} \tilde{B} T \epsilon + \left(\tilde{B} + K\right) \phi(\epsilon)$*. With probability* $1 - \delta_1 - \delta_2$*,*

$$\sup_{\theta \in \Theta_K} |\mathcal{L}_{\text{out}}(\theta) - \mathcal{L}_{\text{in}}(\theta)| = \tilde{\mathcal{O}}\left(\min\left\{\sqrt{\frac{\rho^2 \mathcal{C}(\epsilon) + \log \dfrac{1}{\delta_2}}{NT}}, \frac{\rho \mathcal{C}(\epsilon)^{3/2} + \log \dfrac{1}{\delta_2}}{N\sqrt{T}} + B\epsilon_{\text{app}}\right\}\right).$$

For fast-mixing processes, $\rho$ scales mildly with $T$, e.g., logarithmically for Markov chains with finite mixing time, and Theorem 3.9 recovers the i.i.d. dependency in $N$ as in Theorem 3.5 with an improvement over the $T$ dependency. For non-mixing processes, $\rho$ might scale with $T$ as $\phi(\epsilon)$ is diverging for $\epsilon < 1$ and the rate is vacuous. We work with a relaxation of mixing in Section C.

# 4 Learning In-context Tasks

Assume that there is a set of tasks $\mathcal{W}$ and a prior distribution $\mathcal{P}_{\mathcal{W}}$ over the tasks. Given a task $w \in \mathcal{W}$, there is a prompt distribution $\pi^\star_w$ and a conditional distribution $p^\star_w$ that generates data autoregressively as follows $\vec{z} \sim \pi^\star_w\,, \quad x_t \sim p^\star_w(\cdot \mid \vec{x}_t)$. The task is to learn a model $p_{\hat{\theta}} \in \mathcal{P}$ given access to

$$\forall n \in [N]: \;\; w_n \sim \mathcal{P}_{\mathcal{W}}\,, \;\; \vec{z}^{(n)} \overset{\text{i.i.d}}{\sim} \pi^\star_{w_n}\,, \quad \forall n \in [N], \, t \in [T]: \;\; x^{(n)}_t \sim p^\star_{w_n}(\cdot \mid \vec{x}^{(n)}_t)\,.$$

We denote training, in-sample, in-prompt and out-of-sample losses for a particular task $n$ by $\mathcal{L}^{(n)}_{\text{train}}$, $\mathcal{L}^{(n)}_{\text{in}}$ and $\mathcal{L}^{(n)}_{\text{out}}$, respectively. By $\hat{\mathcal{L}}_{\text{train}}$, $\hat{\mathcal{L}}_{\text{in}}$ and $\hat{\mathcal{L}}_{\text{out}}$, we denote the average losses over all tasks sampled. Similar to Section 2, the learner minimizes $\hat{\mathcal{L}}_{\text{train}}$.

At first, this setting may appear more complex than the standard autoregressive setting, as the learner must learn over a set of autoregressive tasks. However, it can be viewed as a special case of the autoregressive setting. Define $\pi^\star_\infty, p^\star_\infty$ as follows:

$$\pi^\star_\infty(\vec{z}) = \int_{w \in \mathcal{W}} \pi^\star_w(\vec{z}) \mathcal{P}_{\mathcal{W}}(w) dw\,, \quad p^\star_\infty(x \mid \vec{x}_t) = \int_{w \in \mathcal{W}} p^\star_w(x \mid \vec{x}_t)\, \mathcal{P}_{\mathcal{W}}(w \mid \vec{x}_t)\, dw\,, \qquad (2)$$

where $\mathcal{P}_{\mathcal{W}}(w \mid \vec{x}_t) \propto \mathcal{P}_{\mathcal{W}}(w)\, \mathcal{P}_w(\vec{x}_t)$ with $\mathcal{P}_w(\vec{x}_t) := \pi^\star_w(\vec{z}) \prod_{i=1}^{t-1} p^\star_w(x_i \mid \vec{x}_i)$. The conditional distribution $p^\star_\infty$ captures optimal Bayesian inference over sequences. Up to differences in the sampling scheme, we can now interpret the in-context learning as a special case of the autoregressive setting, in which the learner observes samples from the mixture $p^\star_\infty$ and prior $\pi^\star_\infty$. Therefore, it follows that Theorems 3.4 and 3.5 directly extend to the in-context learning setting with $\pi^\star := \pi^\star_\infty, p^\star := p^\star_\infty$ and $\mathcal{L}_{\text{in}}(\theta) := \mathbb{E}[\hat{\mathcal{L}}_{\text{in}}(\theta)], \mathcal{L}_{\text{out}}(\theta) := \mathbb{E}[\hat{\mathcal{L}}_{\text{out}}(\theta)]$.

What makes this setting interesting is that $p^\star_\infty$ is a complicated mixture autogressive process, possibly not satisfying the mixing properties of the individual tasks $p^\star_w$. Nevertheless, the mixing properties of the individual tasks $p^\star_w$ are sufficient to derive generalization bounds for the mixture distribution $p^\star_N$ that approximates $p^\star_\infty$ with the training tasks, i.e., $\mathcal{P}_{\mathcal{W}}$ replaced by $\hat{\mathcal{P}}_{\mathcal{W}} := \sum_{i=1}^N \delta_{w_i}/N$. In the following, set $\pi^\star := \pi^\star_N, p^\star := p^\star_N$.

**Theorem 4.1** (Task-wise out-of-sample generalization)**.** *Let Assumptions 3.1 and 3.7 hold. Let* $\rho$ *be as in Theorem 3.9 where* $\phi(\epsilon)$ *is replaced by* $\max\lceil\phi_{w_i}(\epsilon)\rceil$*. With probability* $1-\delta_1-\delta_2$*,*

$$\sup_{\theta \in \Theta_K} |\hat{\mathcal{L}}_{\text{out}}(\theta) - \hat{\mathcal{L}}_{\text{in}}(\theta)| = \tilde{\mathcal{O}}\left(\min\left\{\sqrt{\frac{\rho^2 \mathcal{C}(\epsilon) + \log \dfrac{1}{\delta_2}}{NT}}, \frac{\rho \mathcal{C}(\epsilon)^{3/2} + \log \dfrac{1}{\delta_2}}{N\sqrt{T}} + B\epsilon_{\text{app}}\right\}\right).$$

Note that the quantity $\hat{\mathcal{L}}_{\text{out}}$ is precisely the loss defined by the out-of-sample generalization error with ground truth $p^\star := p^\star_N$ in Section 3. This follows directly from the fact that in-context sampling and $p^\star_N$ induce the same joint distribution over the observed data.

Theorem 4.1 is a direct extension of Theorem 3.9 to the in-context learning setting. Remarkably, the rate is the same with the looser assumption of mixing properties for the individual tasks. This suggests that in-context learning exploits the task decomposition, allowing for a more efficient learning of the process $p^\star_N$ than Theorem 3.5. Importantly, Theorem 4.1 is not on the ground-truth process $p^\star_\infty$ but on the process $p^\star_N$. We comment more in Section D.

## References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL `https://openreview.net/forum?id=0g0X4H8yN4I`.

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.

Serge Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Mathematische Annalen*, 97:1–59, 1927.

Blair L. Bilodeau, Dylan J. Foster, and Daniel M. Roy. Tight bounds on minimax regret under logarithmic loss via self-concordance. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 919–929. PMLR, 2020. URL `http://proceedings.mlr.press/v119/bilodeau20a.html`.

Etienne Boursier and Nicolas Flammarion. Simplicity bias and optimization threshold in two-layer relu networks. *CoRR*, abs/2410.02348, 2024. doi: 10.48550/ARXIV.2410.02348. URL `https://doi.org/10.48550/arXiv.2410.02348`.

Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2: 499–526, 2002. URL `https://jmlr.org/papers/v2/bousquet02a.html`.

Richard C Bradley. Basic properties of strong mixing conditions. a survey and some open questions. 2005.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL `https://openreview.net/forum?id=TatRHT_1cK`.

Sagnik Chatterjee, Manuj Mukherjee, and Alhad Sethi. Generalization bounds for dependent data using online-to-batch conversion. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL `https://openreview.net/forum?id=MurWORTaF8`.

Paul Doukhan. Mixing. In *Mixing: Properties and Examples*, pages 15–23. Springer, 1995.

Kacha Dzhaparidze and JH Van Zanten. On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117, 2001.

Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, Eran Malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning markov chains. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL `http://papers.nips.cc/paper_files/paper/2024/hash/75b0edb869e2cd509d64d0e8ff446bc1-Abstract-Conference.html`.

Fabian Falck, Ziyu Wang, and Christopher C. Holmes. Is in-context learning in large language models bayesian? A martingale perspective. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=b1YQ5WKY3w`.

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`.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Shi Fu, Yunwen Lei, Qiong Cao, Xinmei Tian, and Dacheng Tao. Sharper bounds for uniformly stable algorithms with stationary mixing process. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL `https://openreview.net/forum?id=8E5Yazboyh`.

Zixuan Gong, Xiaolin Hu, Huayi Tang, and Yong Liu. Towards auto-regressive next-token prediction: In-context learning emerges from generalization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=gK1rl98VRp`.

Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 572–585, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/05546b0e38ab9175cd905eebcc6ebb76-Abstract.html`.

Michael Hahn and Mark Rofin. Why are sensitive functions hard for transformers? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14973–15008. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.800. URL `https://doi.org/10.18653/v1/2024.acl-long.800`.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin. On the minimax regret of sequential probability assignment via square-root entropy. *CoRR*, abs/2503.17823, 2025. doi: 10.48550/ARXIV.2503.17823. URL `https://doi.org/10.48550/arXiv.2503.17823`.

Nirmit Joshi, Gal Vardi, Adam Block, Surbhi Goel, Zhiyuan Li, Theodor Misiakiewicz, and Nathan Srebro. A theory of learning with autoregressive chain of thought. *CoRR*, abs/2503.07932, 2025. doi: 10.48550/ARXIV.2503.07932. URL `https://doi.org/10.48550/arXiv.2503.07932`.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19565–19594. PMLR, 2023. URL `https://proceedings.mlr.press/v202/li23l.html`.

Sanae Lotfi, Yilun Kuang, Marc Finzi, Brandon Amos, Micah Goldblum, and Andrew Gordon Wilson. Unlocking tokens as data points for generalization bounds on larger language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL `http://papers.nips.cc/paper_files/paper/2024/hash/11715d433f6f8b9106baae0df023deb3-Abstract-Conference.html`.

Ron Meir. Nonparametric time series prediction through adaptive model selection. *Mach. Learn.*, 39(1):5–34, 2000. doi: 10.1023/A:1007602715810. URL `https://doi.org/10.1023/A:1007602715810`.

William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Trans. Assoc. Comput. Linguistics*, 11:531–545, 2023. doi: 10.1162/TACL_A_00562. URL `https://doi.org/10.1162/tacl_a_00562`.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 1097–1104. Curran Associates, Inc., 2008. URL `https://proceedings.neurips.cc/paper/2008/hash/7eacb532570ff6858afd2723755ff790-Abstract.html`.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary phi-mixing and beta-mixing processes. *J. Mach. Learn. Res.*, 11:789–814, 2010. doi: 10.5555/1756006.1756032. URL `https://dl.acm.org/doi/10.5555/1756006.1756032`.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *CoRR*, abs/2209.11895, 2022. doi: 10.48550/ARXIV.2209.11895. URL `https://doi.org/10.48550/arXiv.2209.11895`.

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`.

Core Francisco Park, Ekdeep Singh Lubana, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL `https://openreview.net/forum?id=XgH1wfHSX8`.

Alexander Rakhlin and Karthik Sridharan. Sequential probability assignment with binary alphabets and large classes of experts. *CoRR*, abs/1501.07340, 2015. URL `http://arxiv.org/abs/1501.07340`.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/2e10b2c2e1aa4f8083c37dfe269873f8-Abstract-Conference.html`.

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, *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 439–473. PMLR, 2018. URL `http://proceedings.mlr.press/v75/simchowitz18a.html`.

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.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pages 11–30. Springer, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html`.

Mathukumalli Vidyasagar. *Learning and generalisation: with applications to neural networks*. Springer Science & Business Media, 2013.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=RdJVFCHjUMI`.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Oğuz Kaan Yüksel and Nicolas Flammarion. On the sample complexity of next-token prediction. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Oğuz Kaan Yüksel, Mathieu Even, and Nicolas Flammarion. Long-context linear system identification. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=2TuUXtLGhT`.

# A Organization of the appendix

The appendix is organized as follows,

- In Section B, we review the related work.
- In Section C, we extend our results to rephrasability assumption which is looser than mixing.
- In Section D, we discuss our findings in the context of related work.
- In Section E, we provide the preliminary concentration tools.
- In Section F, we prove Theorem 3.4 and its weaker version without Definition 3.2.
- In Section G, we establish the dependence of the rates on the number of sequences $N$. This section provides concentration analysis over independent sequences assuming sequence-level tail bounds.
- In Section H, we establish the dependence of the rates on the number of tokens per sequence $T$. In particular, we establish the tail bounds needed in Section G by an analysis of non-i.i.d. concentration within a single sequence.
- In Section J, we discuss mixing with rephrasability conditions more in detail.
- In Section K, we discuss the concentration of the training mixture for in-context setting.

# B Related Work

Generalization bounds for non-i.i.d. data were first studied in the context of stationary $\beta$-mixing sequences [Yu, 1994, Meir, 2000, Vidyasagar, 2013]. The main technique is the independent block method, introduced by Bernstein [1927], which reduces the problem to the i.i.d. setting. Later analyses incorporated additional assumptions, such as $\phi$-mixing and algorithmic stability, or extended techniques like Rademacher complexity to the non-i.i.d. setting [Mohri and Rostamizadeh, 2008, 2010, Fu et al., 2023, Chatterjee et al., 2025].

Similar to these previous works, our analysis relies on $\phi$-mixing and stability assumptions, and also uses the independent block technique. Our main technical improvements over this body of work are threefold. First, our analysis handles autoregressive processes without a fixed context size. Second, we apply localization [see, e.g., Wainwright, 2019] to obtain faster rates. We build on self-concordance [Bach, 2010, Bilodeau et al., 2020, Ostrovskii and Bach, 2021] arguments by Yüksel et al. [2025] to first concentrate the empirical loss around the denoising loss, then introduce a star-shaped logit space to localize the problem. Lastly, we borrow a notion of regularity for the data-generating process from Yüksel and Flammarion [2025], which serves as the analog of the stability assumption for the hypothesis class, but applied to the data generation process. This allows us to prove generalization bounds without any explicit mixing assumptions.

Closely related is the recent work of Li et al. [2023], who study the sample complexity of in-context learning in i.i.d. and stable first-order dynamical systems. We provide a more general framework that incorporates autoregressive processes of arbitrary order and also derive sharper rates in well-specified settings. Gong et al. [2025] also analyze the sample complexity of in-context learning but do not rigorously track the mixing dependencies that effectively cancel the token dependency, collapsing their rate to the i.i.d. case. Joshi et al. [2025] studies sample complexity but only on the level of sequences, not individual tokens.

In-context learning has been interpreted as implicit Bayesian inference [Xie et al., 2022], but this view has been challenged by Raventós et al. [2023], Falck et al. [2024]. A key criticism is that such inference yields optimal generalization over training tasks, not over new test tasks from the same distribution. However, transformers trained on a sufficiently large number of in-context tasks often generalize well to unseen tasks, suggesting inference over the true task distribution rather than just the training set. Recent work shows that this shift depends on context length, data diversity, and model size [Raventós et al., 2023, Park et al., 2025].

# C Extension to Rephrasability

In Section 3.2, we assume that the data generation process $p^\star$ has a favorable mixing properties such that it is possible to learn the whole dynamics from a single trajectory after sufficiently many samples. However, these assumptions can be restrictive as autoregressive processes of interest such as language are not necessarily mixing. Therefore, we consider an additional intermediate notion of generalization based on fixed prompts:

$$\mathcal{L}_{\text{ppt}}(\theta) \coloneqq \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}^{(n)}_{\text{ppt}}(\theta)\,, \quad \text{with} \quad \mathcal{L}^{(n)}_{\text{ppt}}(\theta) \coloneqq \mathbb{E}_{\vec{x}^{(n)}_{P+1:P+T}}\left[\mathcal{L}^{(n)}_{\text{in}}(\theta)\right]\,. \tag{3}$$

The difference to $\mathcal{L}_{\text{out}}$ is that the expectation is taken only with respect to the sequence and the prompts are kept fixed. The intuition behind Equation (3) is that after a prompt of length $P$, the task is determined and the process associated with a single task may exhibit regularity. Whereas, the joint process induced by the different tasks may not exhibit such regularity as the task is changing with the input sequence.

Furthermore, learning-with-mixing usually leads to rates that are too pessimistic in practice, as evidenced by the results on *learning-without-mixing* [Simchowitz et al., 2018]. These results demonstrate linear dynamical systems that are non-mixing can be learned without any major deflation in the sample size compared to the i.i.d. setting. To derive a result in a similar direction, we adapt the coupling assumption by Yüksel and Flammarion [2025] to our setting:

**Definition C.1** (Couplings)**.** *Let $\tilde{\mathcal{D}}^k$ be all pairs of sequences with the last token different:*

$$\tilde{\mathcal{D}}^k \coloneqq \{(\vec{x},\vec{y}) \in \mathcal{D}^k \times \mathcal{D}^k \mid \forall i \in [k-1] : x_i = y_i\}\,.$$

*Then, let $\Gamma_{\vec{z}}(\vec{x},\vec{y})$ denote the couplings between the continuations after a shared prompt $\vec{z}$:*

$$\Gamma_{\vec{z}}(\vec{x},\vec{y}) \coloneqq \Gamma\left(\mathcal{P}^{1:T-k}_{\vec{z}\circ\vec{x}}, \mathcal{P}^{1:T-k}_{\vec{z}\circ\vec{y}}\right)\,.$$

**Definition C.2** ($r$-rephrasability)**.** *We call the pair $(\pi^\star, p^\star)$ $r$-rephrasable if for all $\vec{z} \in \operatorname{supp}(\pi)$ and $k \in [T]$, the following holds:*

$$\sup_{\vec{x},\vec{y}\in\tilde{\mathcal{D}}^k} \inf_{\gamma\in\Gamma_{\vec{z}}(\vec{x},\vec{y})} \mathbb{E}_{\vec{\mu},\vec{\nu}\sim\gamma}\left[d_H(\vec{\mu},\vec{\nu})\right] \le r\,.$$

Intuitively, a small $r$-rephrasability states that the trajectories that differ by a single token stay close to each other and can be rephrased to each other with a $r$ number of token changes. Notably, the change is not localized within the first $r$ tokens, which would have implied mixing. This is particularly appealing for the prompt-based setting as after the initial prompt, the process has indexed a latent "task" that is shared within the context, whereas a difference in the initial tokens can lead to a completely different task. Therefore, conditioned on the prompt, we expect the data generation process to not bifurcate to completely different paths. We provide a rate for prompt-based generalization that is based on the $r$-rephrasability assumption:

**Theorem C.3** (In-prompt generalization)**.** *Let Assumptions 3.1 and 3.7 hold. Let $\rho \coloneqq \left(\tilde{B} + K\right) r$. With probability $1-\delta_1-\delta_2$,*

$$\sup_{\theta\in\Theta_K} \left|\mathcal{L}_{\text{ppt}}(\theta) - \mathcal{L}_{\text{in}}(\theta)\right| = \tilde{\mathcal{O}}\left(\min\left\{\sqrt{\frac{\rho^2\mathcal{C}(\epsilon)+\log\frac{1}{\delta_2}}{NT}}, \frac{\rho\mathcal{C}(\epsilon)^{3/2}+\log\frac{1}{\delta_2}}{N\sqrt{T}} + B\epsilon_{\text{app}}\right\}\right)\,.$$

# D Discussion

**In-context generalization.** Raventós et al. [2023] describe a phase transition in transformer pretraining from learning the mixture process $p^\star_N$ to the ground-truth process $p^\star_\infty$. Below a certain task diversity threshold, the model learns $p^\star_N$; as diversity increases, it transitions to learn $p^\star_\infty$. This transition depends on the number of optimization steps and the context length [Boursier and Flammarion, 2024, Park et al., 2025]. However, the theoretical understanding of this transition has remained limited.

We hypothesize that the inductive bias of the function class $\mathcal{F}$ play a key role in this transition. As both the number of tasks $N$ and the context length $T$ increase, learning the mixture process $p^\star_N$ with a fixed-capacity model becomes increasingly difficult. The growing complexity of $p^\star_N$ may force the model to generalize across tasks and contexts, effectively biasing it toward the underlying prior distribution and $p^\star_\infty$. This is particularly relevant for in-context learning with linear regression or Markov chain prediction tasks, where $p^\star_\infty$ is shown to be approximately implementable by transformers and usually have a simple structure, e.g., the counting estimator [Akyürek et al., 2023, Edelman et al., 2024].

Our results explain the statistical aspects of in-context generalization. Assuming that optimization error $\epsilon_{\text{opt}}$ is small and the function class is near well-specified, $\epsilon_{\text{app}} \approx 0$, Theorems 3.4 and 4.1 show that the training mixture $p^\star_N$ is learned with a rate that decays with the number of tokens $NT$. If the function class is not well-specified, we obtain a competitive bound. That is, $\hat{\theta}$ is competitive with the best within the class at generalizing to the sequences generated by $p^\star_N$. On the other hand, Theorem 3.5 shows how fast $p^\star_\infty$ is learned with a rate that only depends on $N$, without depending on $T$. This is due to the fact that $\hat{\mathcal{L}}_{\text{out}}(\theta)$ converges around $\mathcal{L}_{\text{out}}(\theta) \coloneqq \mathbb{E}[\hat{\mathcal{L}}_{\text{out}}(\theta)]$, and $p^\star_N$ converges to $p^\star_\infty$ as $N \to \infty$. Thus, for large $N$, learning $p^\star_N$ is equivalent to learning $p^\star_\infty$. We comment more on the speed of convergence of $p^\star_N$ to $p^\star_\infty$ in Section K.

As per the discussion in previous paragraph, when optimization is done well and approximation of $p^\star_N$ is possible, we learn $p^\star_N$ and not $p^\star_\infty$ for finite $N$. Hence, to obtain a rate for $p^\star_\infty$ that decays with $T$, we have two alternatives: either there is an optimization bias such that $\epsilon_{\text{opt}}$ is large, or, there is an inductive bias in the class, i.e., $p^\star_N$ is best approximated by $p^\star_\infty$ within $\mathcal{F}$ as $N$ grows. For the former, our results are not applicable as they are conditioned on a small optimization error. For the latter, our competitive bounds explain why it is possible to generalize to $p^\star_\infty$ with a rate that decays with $NT$. Therefore, we postulate that the inductive bias of the function class is key to understanding generalization of in-context learning. That is, we conjecture that not only growing $N$ but also growing $T$ for a fixed $N$ induces an inductive bias towards $p^\star_\infty$ instead of $p^\star_N$.

**Well-specification.** The results in Sections 3 and 4 show different decay rates of generalization error in $N$. Granted Definition 3.2 with a small constant $\epsilon_{\text{app}}$, the rates are faster in $N$. In order to obtain these faster rates, we develop a star-shaped function class around and use localization techniques. This strategy does not incur any major additional complexity in the sequential entropy of the function class. In contrast, the within-sequence concentration relies on trajectory-dependent properties of the generative model where the martingale variance is difficult to control, even under well-specification, limiting the achievable rate in $T$.

However, when Definition 3.2 holds with a large constant $\epsilon_{\text{app}}$, these bounds may become vacuous. This is when slower rates of order $1/\sqrt{NT}$, matching with the rates of Li et al. [2023], are tighter. Unlike their work, however, we study more general in-context learning settings and do not impose restrictions on the target tasks, such as an i.i.d. structure or first-order stable dynamical systems.

**Stability.** The $K$-stability in Definition 3.6, closely related to the point-wise stability of Li et al. [2023], is key to extending Theorem 3.5 to Theorems C.3 and 3.9. Specifically, point-wise stability of $\frac{K}{m}$, where $m$ is the context size, yields a cumulative bound of $\sum_{m\in[T]} \frac{K}{m} \sim K \log T$ for any input. Li et al. [2023] show that transformers exhibit such stability, with the stability error scaling $\frac{1}{m}$ as the context grows, which implies that our results apply in this setting with at most a logarithmic slowdown in the full sequence length. In addition, they empirically show instances of in-context learning that exhibit stability.

A limitation of such stability bounds are potentially large constants $K$, which can be exponential in the number of transformer layers, as shown by Li et al. [2023]. This stems from the "worst-case" nature of such stability assumptions, being uniform over all contexts and all functions in the class. Similarly, we require the learned estimator to be stable with high probability instead of these worst-case bounds. In practice, such stability may arise from the specific learning algorithm used, such as gradient descent.

Lastly, transformers trained with gradient descent often fail to learn long-range dependencies that violate stability conditions, such as bit-string parity [Hahn and Rofin, 2024], highlighting fundamental computational limitations [Merrill and Sabharwal, 2023]. This suggests that some form of stability, perhaps relaxed, is essential for generalization theory.

**Mixing and rephrasability.** Our convergence rates depend on the regularity of the generative model, as captured by either the $\phi$–mixing constant (Definition 3.8) or the $r$–rephrasability constant (Definition C.2). Both quantities reflect the number of token-level edits required to transform one sequence into another. The $r$-rephrasability is less stringent, since it considers only the total edit count, whereas $\phi$–mixing additionally requires that those edits occur within a fixed window after the two sequences diverge.

We argue that $r$–rephrasability is a more natural condition for language-modeling tasks, since it formalizes the intuitive idea that two sequences differing by a single token can be converted into each other with only a few edits. Moreover, we expect $r$ to decrease as the prompt length $P$ increases, because a longer prompt serves as a richer shared context, making rephrasing easier. Finally, this approach can be extended to structured prompts and to generalization bounds beyond the fixed-prompt setting, which we defer to future work.

We present algorithmic tasks that fit in to the $r$-rephrasability framework that do not mix in Section J. We also sketch how $r$-rephrasability can be estimated.

**Autoregressive settings.** Our autoregressive processes are over a discrete space, reflecting the standard setting in practice. While our results are tailor for discrete sequences, they can be extended to other autoregressive scenarios, such as continuous outputs or regression tasks under mixing conditions. For example, in the continuous case, the model may predict the mean of a Gaussian distribution over outputs, corresponding to the standard squared loss. A natural application is in-context learning with i.i.d. regression observations.

**Other generalization measures.** We focus on sequential metric entropy as our complexity measure. Although chaining methods can yield rates with improved logarithmic dependencies [see, e.g., Wainwright, 2019], we do not pursue them here. A promising direction for future work is to extend our results to other complexity measures, such as Rademacher complexity, as explored by Mohri and Rostamizadeh [2008] for dependent variables.

# E Preliminaries

We introduce additional notation for brevity. The elements of $\mathcal{D}$ are viewed as one-hot vectors of length $d$ with the $i$-th entry equal to 1 if the token is $t_i$ and 0 otherwise. log denotes the natural logarithm.

## E.1 Concentration inequalities

We use concentration inequalities to control tail deviations of empirical processes from their mean. The first result is due to Hoeffding in i.i.d. case [Hoeffding, 1963], extended to the sub-Gaussian case [Wainwright, 2019, Proposition 2.5.]:

**Theorem E.1** (Hoeffding bound). *Let $X_1, \ldots, X_n$ be variables that are independent with mean $\mu_i$ and sub-Gaussian parameter $\sigma_i$. Then, for all $\epsilon \geq 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^{n}\sigma_i^2}\right).$$

We need marginale versions of the Hoeffding bound. Azuma-Hoeffding inequality [Azuma, 1967] gives a concentration result when differences are bounded:

**Theorem E.2** (Azuma-Hoeffding). *Let $Y_0, Y_1, \ldots, Y_n$ be a real-valued martingale sequence that is adapted to the filtration $\mathcal{F}_1, \ldots, \mathcal{F}_n$ where $Y_0 = 0$. Let $X_1, \ldots, X_n$ be the martingale difference sequence, i.e., $X_i = Y_i - Y_{i-1}$. The Azuma-Hoeffding inequality [Azuma, 1967] states that if $|X_i| \le B$ for all $i$, then for any $\epsilon > 0$:*

$$\mathbb{P}\left(\exists k \in [n] : |Y_k| \ge \epsilon\right) \le 2 \exp\left(-\frac{\epsilon^2}{2nB^2}\right) .$$

An improvement of the Azuma-Hoeffding inequality is the Freedman's inequality [Freedman, 1975], which gives a sharper bound when the martingale difference are bounded and has bounded variance:

**Theorem E.3** (Freedman's inequality). *Assume the setting of Theorem E.2. Let $Z_i$ be the conditional variance of $X_i$ given $\mathcal{F}_{i-1}$, i.e., $Z_i = \mathbb{E}[X_i^2 \mid \mathcal{F}_{i-1}]$. Let $W_i = \sum_{j=1}^{i} Z_j$ be the cumulative variance. The Freedman's inequality [Freedman, 1975] states that if $|X_i| \le B$ for all $i$, then for any $\epsilon > 0$ and $W > 0$:*

$$\mathbb{P}\left(\exists k \in [n] : |Y_k| \ge \epsilon \text{ and } W_k \le W\right) \le 2 \exp\left(-\frac{\epsilon^2}{2B\epsilon + 2W}\right) .$$

Lastly, we need an extension of the Freedman's inequality to the case where the martingale difference sequence is not bounded:

**Theorem E.4** (Freedman's inequality with non-bounded differences). *Assume the setting of Theorem E.3. Let $W_i^R$ be the cumulative variance of the martingale difference sequence $X_i$, augmented by the contribution of extremal values beyond $R$, i.e.,*

$$W_i^R = \sum_{j=1}^{i} Z_j + \mathbb{1}_{\{|X_j| > R\}} X_j^2 .$$

*Then, Dzhaparidze and Van Zanten [2001] proves the following for any $\epsilon > 0$ and $W > 0$:*

$$\mathbb{P}\left(\exists k \in [n] : |Y_k| \ge \epsilon \text{ and } W_k^R \le W\right) \le 2 \exp\left(-\frac{3\epsilon^2}{2(R\epsilon + 3W)}\right) .$$

### E.2 Discretization

We use discretization arguments to derive tail bounds of supremum over the function class:

**Proposition E.5** (Discretization). *Assume that $X(\theta) \coloneqq X(f_\theta)$ and $Y(\theta) \coloneqq Y(f_\theta)$ are random variables for any $\theta \in \Theta$ and that we have the following pointwise bound some choice of $\epsilon > 0, \delta > 0$:*

$$\mathbb{P}\left(|X(\theta) - Y(\theta)| \ge \epsilon\right) \le \delta \exp\left(-\mathcal{H}_{2,\infty}\left(\mathcal{F}, \epsilon, T\right)\right) .$$

*Then, we have the supremum bound by discretization:*

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |X(\theta) - Y(\theta)| \ge (1 + 2L)\epsilon\right) \le \delta ,$$

*where $L$ is the Lipschitz constant of $X(f_\theta)$ and $Y(f_\theta)$ with respect to $f_\theta$.*

**Remark E.6.** *All of the losses used in the paper are $\sqrt{2}$-Lipschitz in $\ell_2$ norm with respect to logits $f_\theta$.*

### E.3 Propositions

**Proposition E.7** (Bounded loss). *Given Assumption 3.1, the following holds for all $\theta \in \Theta$:*

$$\sup_{\vec{z} \in \operatorname{supp}(\pi^\star)} \sup_{\vec{x} \in \mathcal{D}^\star} \sup_{x \in \mathcal{D}} -\log p_\theta(x \mid \vec{z} \circ \vec{x}) \le \tilde{B} ,$$

*where $\tilde{B} \coloneqq B + \log d$.*

**Proposition E.8** (Lipschitzness). *Assume there exists a constant $L$ such that for all $\theta, \theta' \in \Theta$:*

$$\sup_{\vec{z} \in \operatorname{supp}(\pi^\star)} \sup_{\vec{x} \in \mathcal{D}^\star} \|f_\theta(\vec{z} \circ \vec{x}) - f_{\theta'}(\vec{z} \circ \vec{x})\|_2 \le L \|\theta - \theta'\|_q .$$

*Then, $\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) \le \mathcal{H}_q(\Theta, \epsilon/L)$.*

# F In-sample generalization

Let $\hat{\theta}$ be any estimate that verifies optimality condition, i.e., $\forall \theta_0 \in \Theta$:

$$\mathcal{L}_{\text{train}}(\hat{\theta}) - \mathcal{L}_{\text{train}}(\theta_0) \leq \epsilon_{\text{opt}} \,. \tag{4}$$

We first prove the slow rate that decays with square root of $NT$ without Definition 3.2:

**Theorem F.1** (Slow in-sample generalization)**.** *Let Assumption 3.1 holds. For any $0 < \delta < 1$, with probability at least $1 - \delta$, the following holds:*

$$\mathcal{L}_{\text{in}}(\hat{\theta}) - \inf_{\theta \in \Theta} \mathcal{L}_{\text{in}}(\theta) = \mathcal{O}\left( \sqrt{\frac{\tilde{B}\left(\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) + \log \frac{e}{\delta}\right)}{NT}} + \epsilon_{\text{opt}} \right) ,$$

*where $\epsilon > 0$ is chosen such that*

$$\epsilon \asymp \sqrt{\frac{\tilde{B}\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) + \log \frac{e}{\delta}}{NT}} \,. \tag{5}$$

*Proof.* A classical strategy to upper bound generalization error is based on the following decomposition:

$$\mathcal{L}_{\text{in}}(\theta) - \mathcal{L}_{\text{in}}(\theta_0) = (\mathcal{L}_{\text{in}}(\theta) - \mathcal{L}_{\text{train}}(\theta)) + (\mathcal{L}_{\text{train}}(\theta) - \mathcal{L}_{\text{train}}(\theta_0)) + (\mathcal{L}_{\text{train}}(\theta_0) - \mathcal{L}_{\text{in}}(\theta_0)) \,.$$

By, Equation (4), we have $\forall \theta_0 \in \Theta$:

$$\mathcal{L}_{\text{in}}(\hat{\theta}) - \mathcal{L}_{\text{in}}(\theta_0) \leq \epsilon_{\text{opt}} + 2 \sup_{\theta \in \Theta} |\mathcal{L}_{\text{in}}(\theta) - \mathcal{L}_{\text{train}}(\theta)| \,. \tag{6}$$

A uniform concentration of $\mathcal{L}_{\text{train}}$ on $\mathcal{L}_{\text{in}}$, combined with Equation (6), yields the desired result. We first establish a pointwise concentration bound for the training loss:

$$\mathcal{L}_{\text{train}}(\theta) - \mathcal{L}_{\text{in}}(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T} - \log p_\theta(x_t^{(n)} \mid \vec{x}_t^{(n)}) + \sum_{n=1}^{N} \sum_{t=1}^{T} \sum_{i=1}^{d} p^\star(e_i \mid \vec{x}_t^{(n)}) \log p_\theta(e_i \mid \vec{x}_t^{(n)}) \,. \tag{7}$$

Since, $x_t^{(n)} \sim p^\star(\cdot \mid \vec{x}_t^{(n)})$, we have that

$$\mathbb{E}_{x_t^{(n)}} \left[ - \log p_\theta(x_t^{(n)} \mid \vec{x}_t^{(n)}) \right] = \sum_{i=1}^{d} p^\star(e_i \mid \vec{x}_t^{(n)}) \log p_\theta(e_i \mid \vec{x}_t^{(n)}) \,.$$

That is, each of the entries of Equation (7) is a martingale difference sequence. Furthermore, Proposition E.7 implies that the martingale difference sequence is bounded by $\tilde{B}$. We can then apply the Azuma-Hoeffding inequality [Azuma, 1967] to obtain the following:

$$\mathcal{P}\left(|\mathcal{L}_{\text{train}}(\theta) - \mathcal{L}_{\text{in}}(\theta)| \geq \epsilon\right) \leq 2 \exp\left( - \frac{NT\epsilon^2}{2\tilde{B}^2} \right) \,.$$

By setting $\epsilon = \sqrt{\frac{2\tilde{B}^2\left(\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) + \log \frac{e}{\delta}\right)}{NT}}$, we obtain the following:

$$\mathcal{P}\left(|\mathcal{L}_{\text{train}}(\theta) - \mathcal{L}_{\text{in}}(\theta)| \geq \epsilon\right) \leq \delta \exp\left(-\mathcal{H}_{2,\infty}\left(\mathcal{F}, \epsilon, T\right)\right) \,.$$

By Proposition E.5, we obtain the uniform bound over the class $\mathcal{F}$:

$$\mathcal{P}\left( \sup_{\theta \in \Theta} |\mathcal{L}_{\text{train}}(\theta) - \mathcal{L}_{\text{in}}(\theta)| \geq \left(1 + 2\sqrt{2}\right) \epsilon \right) \leq \delta \,,$$

Therefore, with probability at least $1 - \delta$:

$$\mathcal{L}_{\text{in}}(\hat{\theta}) - \inf_{\theta \in \Theta} \mathcal{L}_{\text{in}}(\theta) = \mathcal{O}\left( \sqrt{\frac{\tilde{B}^2\left(\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) + \log \frac{e}{\delta}\right)}{NT}} + \epsilon_{\text{opt}} \right) \,.$$

□

Next, we show that Definition 3.2 leads to an improved rate:

**Theorem F.2** (Fast in-sample generalization)**.** *Let Assumption 3.1 and definition 3.2 hold. For any $0 < \delta < 1$, with probability at least $1 - \delta$, the following holds:*

$$\mathcal{L}_{\text{in}}(\hat{\theta}) = \tilde{\mathcal{O}}\left( \frac{B\left(\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) + \log\dfrac{e}{\delta}\right)}{NT} + B\epsilon_{\text{app}} + \epsilon_{\text{opt}} \right) ,$$

*where $\epsilon$ is chosen such that:*

$$\epsilon \asymp \frac{B\left(\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) + \log\dfrac{e}{\delta}\right)}{NT} .$$

*Proof.* Let $h_{\theta_\star}(\vec{x}_t^{(n)}), h^\star(\vec{x}_t^{(n)})$ denote the Hessian of the log-likelihood of the model $p_{\theta_\star}(\cdot \mid \vec{x}_t^{(n)})$ and the target $p^\star(\cdot \mid \vec{x}_t^{(n)})$ with respect to their probabilities, respectively:

$$\begin{aligned} h_{\theta_\star}(\vec{x}_t^{(n)}) &= \operatorname{diag}\left(p_{\theta_\star}(\cdot \mid \vec{x}_t^{(n)})\right) - p_{\theta_\star}(\cdot \mid \vec{x}_t^{(n)}) p_{\theta_\star}(\cdot \mid \vec{x}_t^{(n)})^\top , \\ h^\star(\vec{x}_t^{(n)}) &= \operatorname{diag}\left(p^\star(\cdot \mid \vec{x}_t^{(n)})\right) - p^\star(\cdot \mid \vec{x}_t^{(n)}) p^\star(\cdot \mid \vec{x}_t^{(n)})^\top . \end{aligned}$$

Denote the following quantities:

$$\begin{aligned} \mathcal{Z}_{\text{train}}(\theta) &\coloneqq \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \left\langle p_{\theta^\star}(\cdot \mid \vec{x}_t^{(n)}) - x_t^{(n)}, f_\theta(\vec{x}_t^{(n)}) - f_{\theta^\star}(\vec{x}_t^{(n)}) \right\rangle , \\ \mathcal{Z}^\star_{\text{train}}(\theta) &\coloneqq \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \left\langle p^\star(\cdot \mid \vec{x}_t^{(n)}) - x_t^{(n)}, f_\theta(\vec{x}_t^{(n)}) - f_{\theta^\star}(\vec{x}_t^{(n)}) \right\rangle , \\ \mathcal{V}_{\text{train}}(\theta) &\coloneqq \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \| f_\theta(\vec{x}_t^{(n)}) - f_{\theta^\star}(\vec{x}_t^{(n)}) \|^2_{h_{\theta^\star}(\vec{x}_t^{(n)})} , \\ \mathcal{V}^\star_{\text{train}}(\theta) &\coloneqq \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \| f_\theta(\vec{x}_t^{(n)}) - f_{\theta^\star}(\vec{x}_t^{(n)}) \|^2_{h^\star(\vec{x}_t^{(n)})} . \end{aligned}$$

Note that, by Assumption 3.1 and definition 3.2,

$$\begin{aligned} |\mathcal{Z}_{\text{train}}(\theta) - \mathcal{Z}^\star_{\text{train}}(\theta)| &\le 2B\epsilon_{\text{app}} , \\ |\mathcal{V}_{\text{train}}(\theta) - \mathcal{V}^\star_{\text{train}}(\theta)| &\le 8B^2\epsilon_{\text{app}} . \end{aligned}$$

By properties of the KL divergence [Yüksel and Flammarion, 2025, Proposition A.1.], we can write:

$$\begin{aligned} \mathcal{L}_{\text{train}}(\theta) - \mathcal{L}_{\text{train}}(\theta^\star) + \mathcal{Z}_{\text{train}}(\theta) &= \tilde{\mathcal{L}}_{\text{in}}(\theta) , \quad \text{where} \\ \tilde{\mathcal{L}}_{\text{in}}(\theta) &= \frac{1}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \mathrm{KL}(p_{\theta^\star}(\cdot \mid \vec{y}_t^{(n)}) \parallel p_\theta(\cdot \mid \vec{y}_t^{(n)})) . \end{aligned}$$

By self-concordance of the log likelihood with respect to the logits [Yüksel and Flammarion, 2025, Proposition A.5.], we have the following lower bound:

$$\tilde{\mathcal{L}}_{\text{in}}(\theta) \ge \frac{5}{36B} \mathcal{V}_{\text{train}}(\theta) .$$

Therefore, by Equation (4), we conclude that:

$$\epsilon_{\text{opt}} + CB\epsilon_{\text{app}} + \mathcal{Z}^\star_{\text{train}}(\hat{\theta}) \ge \frac{5}{36B} \mathcal{V}^\star_{\text{train}}(\hat{\theta}) ,$$

for some constant $C$. Two different cases arise:

$$\mathcal{Z}^\star_{\text{train}}(\hat{\theta}) \le \epsilon_{\text{opt}} + CB\epsilon_{\text{app}} \quad \text{and} \quad \mathcal{Z}^\star_{\text{train}}(\hat{\theta}) > \epsilon_{\text{opt}} + CB\epsilon_{\text{app}} .$$

The first case leads to the fast rate:

$$\mathcal{Z}^{\star}_{\text{train}}(\hat{\theta}) = \mathcal{O}\left(\epsilon_{\text{opt}} + B\epsilon_{\text{app}}\right) .$$

The second case leads to the following basic inequality:

$$\mathcal{Z}^{\star}_{\text{train}}(\hat{\theta}) \geq \frac{5}{72B}\mathcal{V}^{\star}_{\text{train}}(\hat{\theta}) .$$

This basic inequality is studied by Yüksel and Flammarion [2025, Appendix B] in the context of Markov chains. Their results directly extend to our setting, by replacing their complexity measure with the sequential metric entropy $\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T)$. In particular, with probability $1-\delta$, for all $\theta \in \Theta$:

$$\mathcal{Z}^{\star}_{\text{train}}(\theta) \leq \max\left\{\frac{5}{72B}\mathcal{V}^{\star}_{\text{train}}(\theta), \alpha\right\} , \quad \text{where} \quad \alpha = \tilde{\mathcal{O}}\left(\frac{B\left(\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) + \log\frac{1}{\delta}\right)}{NT}\right) .$$

□

## G Generalization with independent sequences

We first give the corrected definition of the stability and the stability assumptions in the main text:

**Definition G.1** ($K$-stability)**.** *Let $\mathcal{X}$ be the following set of sequences:*

$$\mathcal{X} \coloneqq \left\{(\vec{x}, \vec{y}) \,|\, \vec{x} \in \mathcal{D}^T \,, |\vec{x}| = |\vec{y}| \,, \exists! i \in [|\vec{x}|] : x_i \neq y_i\right\} .$$

*We say that $\theta \in \Theta$ is $K$-stable if:*

$$\sup_{\vec{z} \in \text{supp}(\pi^\star)} \sup_{\vec{x}, \vec{y} \in \tilde{\mathcal{X}}} \sum_{i=1}^{|\vec{x}|} \|f_\theta(\vec{z} \circ \vec{x}_{1:i}) - f_\theta(\vec{z} \circ \vec{y}_{1:i})\|_2 \leq K .$$

**Assumption G.2.** *Let $\Theta_K \coloneqq \{\theta \in \Theta : \theta \text{ is } K-\text{stable}\}$. There exist a $\delta_1 > 0$ such that*

$$\mathbb{P}\left(\{\hat{\theta} \in \Theta_K\} \cap \{\text{argmin}_\theta \, \mathcal{L}(\theta) \in \Theta_K\}\right) \geq 1 - \delta_1 ,$$

*where $\mathcal{L}$ is the generalization loss of interest.*

In this section, we establish how the rates in Sections C, 3.2 and 4 depend on the number of sequences $N$. We provide a unified treatment for all theorems by abstracting away the details of per-sequence concentrations and assume a tail bound on the in-sample loss. These tail bounds are established in Section H for the in-sample loss and tailored to specific settings. Similar to Section F, we establish both a slow and a fast rate in $N$, depending on if Definition 3.2 is granted.

**Stability.** We work conditionally on the event in Assumption G.2. This is why Theorems C.3, 3.9 and 4.1 pay the constant that depends on the probability of this event. In the discussion below, we take it granted that $\hat{\theta} \in \Theta_K$ and $\text{argmin}_\theta \, \mathcal{L}(\theta) \in \Theta_K$.

**Decomposition of loss.** Our generalization results are given in terms of

$$\sup_{\theta \in \Theta_K} |\mathcal{L}(\theta) - \mathcal{L}_{\text{in}}(\theta)| .$$

Recall that this controls the generalization error

$$\mathcal{L}(\hat{\theta}) - \inf_{\theta \in \Theta_K} \mathcal{L}(\theta) = \mathcal{L}(\hat{\theta}) - \inf_{\theta \in \Theta} \mathcal{L}(\theta) ,$$

as the following decomposition relates the in-sample to any other generalization loss for $\theta_0 \in \Theta_K$:

$$\begin{aligned} \mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta_0) &= \left(\mathcal{L}(\hat{\theta}) - \mathcal{L}_{\text{in}}(\hat{\theta})\right) + \left(\mathcal{L}_{\text{in}}(\hat{\theta}) - \mathcal{L}_{\text{in}}(\theta_0)\right) + \left(\mathcal{L}_{\text{in}}(\theta_0) - \mathcal{L}(\theta_0)\right) , \\ &\leq \left(\mathcal{L}_{\text{in}}(\hat{\theta}) - \mathcal{L}_{\text{in}}(\theta_0)\right) + 2\sup_{\theta \in \Theta_K} |\mathcal{L}(\theta) - \mathcal{L}_{\text{in}}(\theta)| . \end{aligned} \tag{8}$$

The first term is bounded in Section F and we work only on the second term.

**Tail bound assumption.** Let $\mathcal{L}^{(n)}(\theta), \mathcal{L}(\theta)$ denote the expected in-sample loss and its average, respectively:

$$\mathcal{L}^{(n)}(\theta) \coloneqq \mathbb{E}\left[\mathcal{L}_{\text{in}}^{(n)}(\theta)\right] , \quad \mathcal{L}(\theta) \coloneqq \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}^{(n)}(\theta) .$$

This expectation can be tailored to specific settings, such as the out-of-sample loss $\mathcal{L}_{\text{out}}$, prompt-based loss $\mathcal{L}_{\text{ppt}}$ and in-context learning loss, which we comment in Section G.3. In the following, we assume the sub-Gaussianity of the in-sample loss $\mathcal{L}_{\text{in}}^{(n)}(\theta)$ where we derive the sub-Gaussianity parameter in Section H:

**Assumption G.3.** *$\mathcal{L}_{\text{in}}^{(n)}(\theta)$ is sub-Gaussian with parameter $\sigma^2$ for all $\theta \in \Theta_K$.*

### G.1 Slow rate.

**Theorem G.4** (Slow rate)**.** *Let Assumptions 3.1 and G.3 hold. With probability $1-\delta$,*

$$\sup_{\theta\in\Theta_K} |\mathcal{L}(\theta) - \mathcal{L}_{\text{in}}(\theta)| = \tilde{\mathcal{O}}\left(\sigma\sqrt{\frac{\mathcal{H}_{2,\infty}(\mathcal{F},\epsilon,T) + \log\frac{e}{\delta}}{N}}\right) ,$$

*where $\epsilon$ is chosen such that:*

$$\epsilon \asymp \sigma\sqrt{\frac{\mathcal{H}_{2,\infty}(\mathcal{F},\epsilon,T) + \log\frac{e}{\delta}}{N}} .$$

By Theorem E.1, we have the following tail bound for any $\epsilon > 0$:

$$\mathbb{P}\left(|\mathcal{L}(\theta) - \mathcal{L}_{\text{in}}(\theta)| \geq \epsilon\right) \leq 2\exp\left(-\frac{N\epsilon^2}{2\sigma^2}\right) .$$

By setting $\epsilon = \sigma\sqrt{\frac{\left(\mathcal{H}_{2,\infty}(\mathcal{F},\epsilon,T) + \log\frac{e}{\delta}\right)}{2N}}$, we obtain

$$\mathbb{P}\left(|\mathcal{L}(\theta) - \mathcal{L}_{\text{in}}(\theta)| \geq \epsilon\right) \leq \delta\exp\left(-\mathcal{H}_{2,\infty}(\mathcal{F},\epsilon,T)\right) .$$

By Proposition E.5, we obtain the uniform bound over the class $\mathcal{F}$:

$$\mathcal{P}\left(\sup_{\theta\in\Theta_K} |\mathcal{L}(\theta) - \mathcal{L}_{\text{in}}(\theta)| \geq \left(1+2\sqrt{2}\right)\epsilon\right) \leq \delta .$$

### G.2 Fast rate for Theorem 3.9.

**Theorem G.5** (Fast rate)**.** *Let Assumptions 3.1 and G.3 and definition 3.2 hold. With probability $1-\delta$,*

$$\sup_{\theta\in\Theta_K} |\mathcal{L}(\theta) - \mathcal{L}_{\text{in}}(\theta)| = \tilde{\mathcal{O}}\left(\max\left\{\frac{\sigma\left(\mathcal{H}_{2,\infty}(\mathcal{F},\epsilon,T) + \log\frac{e}{\delta}\right)\sqrt{\mathcal{H}_{2,\infty}(\mathcal{F},\epsilon,T) + \log N + \log\frac{e}{\delta}}}{N} + B\epsilon_{\text{app}}, \mathcal{L}_{\text{in}}(\theta)\right\}\right) ,$$

*where $\epsilon$ is chosen such that:*

$$\epsilon \asymp \frac{\sigma\left(\mathcal{H}_{2,\infty}(\mathcal{F},\epsilon,T) + \log\frac{e}{\delta}\right)\sqrt{\mathcal{H}_{2,\infty}(\mathcal{F},\epsilon,T) + \log N + \log\frac{e}{\delta}}}{N} .$$

We first enlarge the function class to be star-shaped around the target $p_{\theta^\star}$. This is necessary to obtain the fast decay in the number of tokens. Let $\tilde{\mathrm{KL}} : \Delta^{d-1} \times \mathbb{R}^d \to \mathbb{R}$ be the following function:

$$\tilde{\mathrm{KL}}(p \parallel q) \coloneqq \mathrm{KL}(p \parallel \sigma(q)) \,.$$

$\tilde{\mathrm{KL}}$ is convex and differentiable in the second argument. Therefore, given two vectors $q_1, q_2 \in \mathbb{R}^d$, we have the following:

$$\begin{aligned}&\{\mathrm{KL}(p \parallel \sigma(\lambda q_1 + (1-\lambda) q_2)) \mid \lambda \in [0,1]\}\\ &\qquad \subseteq [\min\{\mathrm{KL}(p \parallel q_1), \mathrm{KL}(p \parallel q_2)\}, \max\{\mathrm{KL}(p \parallel q_1), \mathrm{KL}(p \parallel q_2)\}] \,.\end{aligned}$$

Moreover, for any $\lambda \in [0,1]$, there exist a $\lambda^\star \in [0,1]$ such that

$$\tilde{\mathrm{KL}}(p \parallel \lambda^\star q_1 + (1-\lambda^\star) q_2) = \lambda \tilde{\mathrm{KL}}(p \parallel q_1) + (1-\lambda)\tilde{\mathrm{KL}}(p \parallel q_2) \,,$$

Now, fix any prompt $\vec{z} \in \mathrm{supp}(\pi)$ and context $\vec{x} \in \mathcal{D}^\star$ and $\theta \in \Theta_K$. Let $\vec{y} \coloneqq \vec{z} \circ \vec{x}$ for brevity. For any $\lambda \in [0,1]$, there exist a $\lambda^\star \in [0,1]$ such that

$$\begin{aligned}&\tilde{\mathrm{KL}}(p^\star(\cdot \mid \vec{y}) \parallel \lambda^\star f_\theta(\vec{y}) + (1-\lambda^\star) f_{\theta^\star}(\vec{y}))\\ &\qquad = \lambda \tilde{\mathrm{KL}}(p^\star(\cdot \mid \vec{y}) \parallel f_\theta(\vec{y})) + (1-\lambda)\tilde{\mathrm{KL}}(p^\star(\cdot \mid \vec{y}) \parallel f_{\theta^\star}(\vec{y})) \,,\end{aligned}$$

By using the above property, we define the following function:

$$f_{\theta,\lambda}(\vec{y}) = \lambda^\star f_\theta(\vec{y}) + (1-\lambda^\star) f_{\theta^\star}(\vec{y}) \,.$$

Clearly, we have that

$$\begin{aligned}&\tilde{\mathrm{KL}}(p^\star(\cdot \mid \vec{y}) \parallel f_{\theta,\lambda}(\vec{y})) - \tilde{\mathrm{KL}}(p^\star(\cdot \mid \vec{y}) \parallel f_{\theta^\star}(\vec{y}))\\ &\qquad = \lambda \left(\tilde{\mathrm{KL}}(p^\star(\cdot \mid \vec{y}) \parallel f_\theta(\vec{y})) - \tilde{\mathrm{KL}}(p^\star(\cdot \mid \vec{y}) \parallel f_{\theta^\star}(\vec{y}))\right) \,.\end{aligned}$$

By using the same construction point-wise for each $\vec{y}$ and $\theta \in \Theta_K$, we can define the following function class:

$$\mathcal{F}_{\Theta_K \times [0,1]} = \{f_{\theta,\lambda} \mid f_\theta \in \mathcal{F}_{\Theta_K}\} \,.$$

We extend the definitions of loss functions $\mathcal{L}_{\mathrm{in}}, \mathcal{L}$ to the function class $\Theta_K \times [0,1]$. Let $\Delta_{\mathrm{in}}, \Delta$ be the shifted versions of the loss functions:

$$\begin{aligned}\Delta_{\mathrm{in}}(\theta,\lambda) &\coloneqq \mathcal{L}_{\mathrm{in}}(\theta,\lambda) - \mathcal{L}_{\mathrm{in}}(\theta^\star) \,,\\ \Delta(\theta,\lambda) &\coloneqq \mathcal{L}(\theta,\lambda) - \mathcal{L}(\theta^\star) \,.\end{aligned}$$

We are ready to state the localization lemma:

**Lemma G.6** (Localization)**.** *Fix any $r > 0$. Let $\Theta_r$, $\mathcal{S}_r$ and $\mathcal{B}_r$ denote the following set, supremum and event:*

$$\begin{aligned}\Theta_r &\coloneqq \{(\theta,\lambda) \in \Theta_K \times [0,1] \mid \Delta(\theta,\lambda) \le r\} \,,\\ \mathcal{S}_r &\coloneqq \sup_{(\theta,\lambda)\in\Theta_r} |\Delta(\theta,\lambda) - \Delta_{\mathrm{in}}(\theta,\lambda)| \,,\\ \mathcal{B}_r &\coloneqq \{\forall(\theta,\lambda) \in \Theta_K \times [0,1] : |\Delta(\theta,\lambda) - \Delta_{\mathrm{in}}(\theta,\lambda)| \le \frac{1}{2}|\Delta(\theta,\lambda)| + \frac{r}{2}\} \,.\end{aligned}$$

*Then, we have the following relation:*

$$\mathbb{P}\left(\mathcal{B}_r\right) \ge \mathbb{P}(\mathcal{S}_r \le \frac{r}{2}) \,.$$

*Proof.* We show that the following holds:

$$\mathcal{B}(r)^c \subseteq \{\mathcal{S}_r > \frac{r}{2}\} \,.$$

Assume that there exist a pair $(\theta,\lambda) \in \Theta_K \times [0,1]$ such that

$$|\Delta(\theta,\lambda) - \Delta_{\mathrm{in}}(\theta,\lambda)| > \frac{1}{2}|\Delta(\theta,\lambda)| + \frac{r}{2}$$

We show that this implies that $\mathcal{S}_r > \frac{r}{2}$.

If $\Delta(\theta, \lambda) \leq r$, then we have:

$$\mathcal{S}_r \geq |\Delta(\theta, \lambda) - \Delta_{\mathrm{in}}(\theta, \lambda)| > \frac{r}{2} \, .$$

Otherwise, $\Delta(\theta, \lambda) > r$. Setting $\lambda' \coloneqq \frac{r}{\Delta(\theta, \lambda)} \lambda$, we have

$$\Delta(\theta, \lambda') = r \quad \text{and} \quad |\Delta(\theta, \lambda') - \Delta_{\mathrm{in}}(\theta, \lambda')| > \frac{r}{2} \, .$$

□

**Corollary G.7** (Misspecified localization)**.** *Let Definition 3.2 hold. Using the definitions in Lemma G.6,*

$$\mathbb{P}\left(\forall \theta \in \Theta_K : |\mathcal{L}(\theta) - \mathcal{L}_{\mathrm{in}}(\theta)| \leq \frac{3}{2}\left(r + 3B\epsilon_{\mathrm{app}}\right) \, , \;\; \textit{or} \;\; |\mathcal{L}(\theta) - \mathcal{L}_{\mathrm{in}}(\theta)| \leq 3\mathcal{L}_{\mathrm{in}}(\theta)\right) \geq \mathbb{P}(\mathcal{S}_r \leq \frac{r}{2}) \, .$$

*Proof.* Assume that the event $\mathcal{S}_r \leq \frac{r}{2}$ holds. By Lemma G.6, we have that $\mathcal{B}_r$ holds. Then, we have the following by Definition 3.2:

$$\begin{aligned} |\mathcal{L}(\theta) - \mathcal{L}_{\mathrm{in}}(\theta)| &\leq |\mathcal{L}(\theta^\star) - \mathcal{L}_{\mathrm{in}}(\theta^\star)| + |\Delta(\theta, 0) - \Delta_{\mathrm{in}}(\theta, 0)| \\ &\leq \frac{1}{2}|\Delta(\theta, 0)| + \frac{r}{2} + B\epsilon_{\mathrm{app}} \\ &\leq \frac{1}{2}\mathcal{L}(\theta) + \frac{r}{2} + \frac{3B\epsilon_{\mathrm{app}}}{2} \, . \end{aligned}$$

If $\mathcal{L}(\theta) \leq 2\left(r + 3B\epsilon_{\mathrm{app}}\right)$, then we have

$$|\mathcal{L}(\theta) - \mathcal{L}_{\mathrm{in}}(\theta)| \leq \left(r + 3B\epsilon_{\mathrm{app}}\right) + \frac{r}{2} + \frac{3B\epsilon_{\mathrm{app}}}{2} = \frac{3}{2}\left(r + 3B\epsilon_{\mathrm{app}}\right) \, .$$

If $\mathcal{L}(\theta) > 2\left(r + 3B\epsilon_{\mathrm{app}}\right)$, then we have

$$|\mathcal{L}(\theta) - \mathcal{L}_{\mathrm{in}}(\theta)| \leq \frac{3}{4}\mathcal{L}(\theta) \, .$$

In this case, we have that $\mathcal{L}(\theta) \leq 4\mathcal{L}_{\mathrm{in}}(\theta)$. □

According to Lemma G.6 and Corollary G.7, we just need to lower bound $\mathbb{P}(\mathcal{S}_r \leq \frac{r}{2})$ for a suitable choice of $r$. We replace $\mathcal{S}_r \leq \frac{r}{2}$ with the event $\tilde{\mathcal{S}}_r \leq \frac{r}{5}$ which is defined as follows:

$$\begin{aligned} \tilde{\mathcal{S}}_r &\coloneqq \sup_{(\theta, \lambda) \in \tilde{\Theta}_r} |\mathcal{L}(\theta, \lambda) - \mathcal{L}_{\mathrm{in}}(\theta, \lambda)| \, , \quad \text{where} \\ \tilde{\Theta}_r &\coloneqq \{(\theta, \lambda) \in \Theta_K \times [0, 1] \mid \mathcal{L}(\theta, \lambda) \leq r\} \, . \end{aligned}$$

Note that $\mathbb{P}\left(\tilde{\mathcal{S}}_r \leq \frac{r}{5}\right)$ is increasing in $r$ and $\tilde{\mathcal{S}}_{r + B\epsilon_{\mathrm{app}}} \leq \frac{r}{2} - B\epsilon_{\mathrm{app}}$ implies that $\mathcal{S}_r \leq \frac{r}{2}$ holds. For a suitable choice of $r \geq 4B\epsilon_{\mathrm{app}}$, we have that

$$\begin{aligned} \mathbb{P}\left(\mathcal{S}_r \leq \frac{r}{2}\right) &\geq \mathbb{P}\left(\tilde{\mathcal{S}}_{r + B\epsilon_{\mathrm{app}}} \leq \frac{r}{2} - B\epsilon_{\mathrm{app}}\right) \\ &\geq \mathbb{P}\left(\tilde{\mathcal{S}}_{r + B\epsilon_{\mathrm{app}}} \leq \frac{r + B\epsilon_{\mathrm{app}}}{5}\right) \\ &\geq \mathbb{P}\left(\tilde{\mathcal{S}}_r \leq \frac{r}{5}\right) \, . \end{aligned}$$

For $r \leq 4B\epsilon_{\mathrm{app}}$, we obtain the desired rate by resetting $r = 4B\epsilon_{\mathrm{app}}$.

We start by first establishing a point-wise bound. Consider the following decomposition:

$$\mathcal{L}(\theta, \lambda) - \mathcal{L}_{\text{in}}(\theta, \lambda) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}^{(n)}(\theta, \lambda) - \mathcal{L}_{\text{in}}^{(n)}(\theta, \lambda) \,.$$

This is a martingale sequence as each $d^{(n)}(\theta, \lambda) := \mathcal{L}^{(n)}(\theta, \lambda) - \mathcal{L}_{\text{in}}^{(n)}(\theta, \lambda)$ has zero-mean. Let $W_N^R(\theta, \lambda)$ be the quadratic variation

$$W_N^R(\theta, \lambda) := \frac{1}{N^2} \sum_{n=1}^{N} \mathbb{E}\left[d^{(n)}(\theta, \lambda)^2\right] + \mathbb{1}_{\{\frac{1}{N}|d^{(n)}(\theta,\lambda)|>R\}} d^{(n)}(\theta, \lambda)^2 \,.$$

By Theorem E.4, we obtain

$$\mathbb{P}\left(|\mathcal{L}(\theta, \lambda) - \mathcal{L}_{\text{in}}(\theta, \lambda)| \geq \epsilon \;\text{ and }\; W_N^R(\theta, \lambda) \leq W\right) \leq 2 \exp\left(-\frac{3\epsilon^2}{2(R\epsilon + 3W)}\right) .$$

Setting $\epsilon = c_1 \dfrac{W}{R}$ for some constant $c_1 > 0$,

$$\mathbb{P}\left(|\mathcal{L}(\theta, \lambda) - \mathcal{L}_{\text{in}}(\theta, \lambda)| \geq c_1 \frac{W}{R}\right) \leq 2 \exp\left(-c_2 \frac{W}{R^2}\right) + \mathbb{P}\left(W_N^R(\theta, \lambda) > W\right) ,$$

for some constant $c_2 > 0$. Let $\mathcal{N}_\epsilon$ be an $\epsilon$-net of $\mathcal{F}|_{U_T}$. Then, by the argument in Proposition E.5,

$$\mathbb{P}\left(\sup_{\theta \in \Theta_r} |\mathcal{L}(\theta, \lambda) - \mathcal{L}_{\text{in}}(\theta, \lambda)| \geq c_1(1 + 2\sqrt{2})\frac{W}{R}\right) \leq 2 \exp\left(-c_2 \frac{W}{R^2} + \mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T)\right) + \sum_{\theta \in \mathcal{N}_\epsilon} \mathbb{P}\left(W_N^R(\theta, \lambda) > W\right) .$$

The second term can be bounded by ensuring that $\forall \theta \in \mathcal{N}_\epsilon$:

$$\mathbb{P}\left(W_N^R(\theta, \lambda) > W\right) \leq \delta_1 \exp\left(-\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T)\right) .$$

Assume that $R$ is such that

$$\mathbb{P}\left(\sup_n d^{(n)}(\theta, \lambda) > NR\right) \leq \delta_1 \exp\left(-\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T)\right) .$$

Then, we have the following conditional on the event $\{\sup_n d^{(n)}(\theta, \lambda) \leq NR\}$:

$$W_N^R(\theta, \lambda) = \frac{1}{N^2} \sum_{n=1}^{N} \mathbb{E}\left[d^{(n)}(\theta, \lambda)^2\right] \leq \frac{R}{N} \sum_{n=1}^{N} \mathbb{E}\left[d^{(n)}(\theta, \lambda)\right] = Rr \,.$$

Therefore, we can choose $W = Rr$ and $R = \dfrac{\sigma\sqrt{2}\sqrt{\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) + \log N + \log \frac{e}{\delta_1}}}{N}$. The first term can be bounded by ensuring that

$$\epsilon = \frac{c_1}{c_2} R \left(\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) + \log \frac{2}{\delta_2}\right) .$$

Finally, we have the following bound:

$$\mathbb{P}\left(\sup_{\theta \in \Theta_K} |\mathcal{L}(\theta, \lambda) - \mathcal{L}_{\text{in}}(\theta, \lambda)| \geq 5c_1 r\right) \leq \delta_1 + \delta_2 \,.$$

By choosing $\delta_1, \delta_2$ for the desired confidence interval and $c_1 = \dfrac{1}{25}$, we conclude the proof.

### G.3 Proofs of Theorems C.3, 3.9 and 4.1

In all theorems, the concentration with independent sequences rely on the same analysis. Here, we clarify the applications of these results to Theorems C.3, 3.9 and 4.1:

- For out-of-sample generalization, the expectation is taken with respect to the distribution of the sequence $n$ and the prompt $\vec{z}^{(n)}$, leading to $\mathcal{L}^{(n)}(\theta) = \mathcal{L}_{\text{out}}(\theta)$.
- For in-prompt generalization, the expectation is taken with respect to the distribution of the sequence $n$, leading to $\mathcal{L}^{(n)}(\theta) = \mathcal{L}_{\text{ppt}}(\theta)$.
- For in-context generalization, the expectation is taken with respect to the distribution of the sequence $n$ and the prompt $\vec{z}^{(n)}$ for the task $w_n$, leading to $\mathcal{L}^{(n)}(\theta) = \mathcal{L}_{\text{out}}^{(n)}(\theta)$.

We explain the tail bounds for each setting in Section H.

## H Generalization within a sequence

In this section, we establish how the rates in Sections C, 3.2 and 4 depend on the number of tokens $T$ in a single sequence for $\theta \in \Theta_K$. In particular, we establish a tail bound, which shows sub-Gaussian concentration of the in-sample loss $\mathcal{L}_{\text{in}}^{(n)}(\theta)$ around its expectation $\mathbb{E}\left[\mathcal{L}_{\text{in}}^{(n)}(\theta)\right]$. Combined with Section G, this tail bound implies Theorems C.3 and 3.9. The in-context setting of these two results are the same with no difference in the proof as we only study a single task here, reducing to the same problem.

Note that for out-of-sample generalization $\mathcal{L}_{\text{out}}$, the expectation $\mathbb{E}\left[\mathcal{L}_{\text{in}}^{(n)}(\theta)\right]$ is taken over the sequence of tokens and the prompt, whereas for in-prompt generalization $\mathcal{L}_{\text{ppt}}$, it is only taken over the sequence of tokens.

### H.1 Proof of Theorems 3.9 and 4.1

Let $\vec{x}^{(n)} := \left(x_1^{(n)}, \ldots, x_T^{(n)}\right)$ be the full token sequence. We divide $\vec{x}^{(n)}$ into parts of size $\phi := \phi(\epsilon)$ for Theorem 3.9 or $\phi := \phi_{\mathcal{W}}(\epsilon)$ for Theorem 4.1. Denote the parts of $\vec{x}^{(n)}$ as $\vec{y}_1^{(n)}, \ldots, \vec{y}_M^{(n)}$. We have the following decomposition of the difference of the in-sample and the in-prompt loss:

$$\mathcal{L}_{\text{in}}^{(n)}(\theta) - \mathcal{L}_{\text{out}}^{(n)}(\theta) = \frac{1}{T}\sum_{m=0}^{M}\left(\mathbb{E}\left[\mathcal{L}_{\text{in}}^{(n)}(\theta) \mid \mathcal{F}_m^{(n)}\right] - \mathbb{E}\left[\mathcal{L}_{\text{in}}^{(n)}(\theta) \mid \mathcal{F}_{m-1}^{(n)}\right]\right),$$

where $\mathcal{F}_m^{(n)}$ is the $\sigma$-algebra generated by the first $m$ parts of the sequence and the prompt with the convention $\mathbb{E}\left[\mathcal{L}_{\text{in}}^{(n)}(\theta) \mid \mathcal{F}_{-1}^{(n)}\right] = \mathcal{L}_{\text{out}}^{(n)}(\theta)$. It is clear that this is a martingale sequence as

$$\mathbb{E}\left[\mathbb{E}\left[\mathcal{L}_{\text{in}}^{(n)}(\theta) \mid \mathcal{F}_m^{(n)}\right] \mid \mathcal{F}_{m-1}^{(n)}\right] = \mathbb{E}\left[\mathcal{L}_{\text{in}}^{(n)}(\theta) \mid \mathcal{F}_{m-1}^{(n)}\right].$$

We bound the differences at each step as follows. Moreover, we have the following by the mixing property:

$$\left\|\mathcal{P}_{\vec{z}^{(n)}}^{m\phi:T}\left(\cdot \mid \mathcal{F}_{m-1}^{(n)}\right) - \mathcal{P}^{m\phi:T}(\cdot)\right\|_{\text{TV}} \le \epsilon, \quad \left\|\mathcal{P}_{\vec{z}^{(n)}}^{(m+1)\phi:T}\left(\cdot \mid \mathcal{F}_m^{(n)}\right) - \mathcal{P}^{(m+1)\phi:T}(\cdot)\right\|_{\text{TV}} \le \epsilon.$$

This implies that

$$\left\|\mathcal{P}_{\vec{z}^{(n)}}^{(m+1)\phi:T}\left(\cdot \mid \mathcal{F}_{m-1}^{(n)}\right) - \mathcal{P}_{\vec{z}^{(n)}}^{(m+1)\phi:T}\left(\cdot \mid \mathcal{F}_m^{(n)}\right)\right\|_{\text{TV}} \le 2\epsilon, \tag{9}$$

and that the difference is bounded by

$$\left|\mathbb{E}\left[\mathcal{L}_{\text{in}}^{(n)}(\theta) \mid \mathcal{F}_m^{(n)}\right] - \mathbb{E}\left[\mathcal{L}_{\text{in}}^{(n)}(\theta) \mid \mathcal{F}_{m-1}^{(n)}\right]\right| \le \underbrace{2\epsilon\tilde{B}\left(T - (m+1)\phi\right)}_{(A)} + \underbrace{2\phi\left(\tilde{B} + \sqrt{2}K\right)}_{(B)}.$$

The first term $(A)$ upper bounds the difference in loss by the maximum loss $\tilde{B} \times (T - (m+1)\phi)$ when the conditional distributions do not match. The second term $(B)$ upper bounds the difference when the sequences generated match in indices between $(m+1)\phi$ and $T$. In this case, there are at most $2\phi$ tokens in indices between $(m-1)\phi$ and $(m+1)\phi$ that differ in the two continuations. This results at most $2\phi\tilde{B}$ loss for these first $2\phi$ tokens. By Definition G.1, the differences in predictions in the logit space for positions between $(m-1)\phi$ and $T$ are bounded by $2\phi \times K$ in $\|\cdot\|_2$ norm. Since, the loss is $\sqrt{2}$-Lipschitz with respect to $\|\cdot\|_2$ norm in the logit space, the loss is bounded by $2\phi\left(\tilde{B} + \sqrt{2}K\right)$.

We set $\rho$ by minimizing the following upper bound on the differences:

$$\rho := \inf_{\epsilon \in (0,1)} 2\epsilon\tilde{B}\left(T - (m+1)\phi\right) + 2\phi(\epsilon)\left(\tilde{B} + \sqrt{2}K\right) .$$

Now, we have a martingale difference sequence with $\frac{\rho}{T}$-bounded differences. By Theorem E.2, we have the following for any $\epsilon > 0$:

$$\mathbb{P}\left(\left|\mathcal{L}^{(n)}_{\text{in}}(\theta) - \mathcal{L}^{(n)}_{\text{out}}(\theta)\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{T\epsilon^2}{2\rho^2}\right) .$$

Therefore, sub-Gaussian concentration holds with $\sigma^2 = \mathcal{O}\left(\frac{\rho}{\sqrt{T}}\right)$.

### H.2 Proof of Theorem C.3

We use a similar decomposition of the in-sample loss $\mathcal{L}^{(n)}_{\text{in}}(\theta)$ as in previous section:

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

where we use blocks of single-tokens instead of $\phi$. Now, first note that

$$\left|\mathbb{E}\left[\mathcal{L}^{(n)}_{\text{in}}(\theta) \mid \mathcal{F}^{(n)}_t\right] - \mathbb{E}\left[\mathcal{L}^{(n)}_{\text{in}}(\theta) \mid \mathcal{F}^{(n)}_{t-1}\right]\right| \leq \sup_{\tilde{\mathcal{F}}^{(n)}_t}\left|\mathbb{E}\left[\mathcal{L}^{(n)}_{\text{in}}(\theta) \mid \mathcal{F}^{(n)}_t\right] - \mathbb{E}\left[\mathcal{L}^{(n)}_{\text{in}}(\theta) \mid \tilde{\mathcal{F}}_t^{(n)}\right]\right| ,$$

where the supremum is taken over all $\sigma$-algebras $\tilde{\mathcal{F}}^{(n)}_t \subset \mathcal{F}^{(n)}_{t-1}$ that are generated by the choice of the token at position $t$. Let $\gamma$ be a coupling between the two distributions $\mathcal{P}^{t:T}_{\vec{z}^{(n)}}\left(\cdot \mid \mathcal{F}^{(n)}_t\right)$ and $\mathcal{P}^{t:T}_{\vec{z}^{(n)}}\left(\cdot \mid \tilde{\mathcal{F}}^{(n)}_t\right)$, i.e.,

$$\gamma \in \Gamma\left(\mathcal{P}^{t:T}_{\vec{z}^{(n)}}\left(\cdot \mid \mathcal{F}^{(n)}_t\right), \mathcal{P}^{t:T}_{\vec{z}^{(n)}}\left(\cdot \mid \tilde{\mathcal{F}}^{(n)}_t\right)\right) .$$

Let $k$ be the following quantity:

$$k := \mathbb{E}_{\vec{\mu},\vec{\nu}\sim\gamma}\left[d_H(\vec{\mu},\vec{\nu})\right] .$$

$k$ quantifies the number of distinct tokens in the two suffixes that complete the original sequence at position $t$, measured by the coupling $\gamma$.

This implies that the difference is bounded by

$$\left|\mathbb{E}\left[\mathcal{L}^{(n)}_{\text{in}}(\theta) \mid \mathcal{F}^{(n)}_t\right] - \mathbb{E}\left[\mathcal{L}^{(n)}_{\text{in}}(\theta) \mid \tilde{\mathcal{F}}^{(n)}_t\right]\right| \leq \underbrace{k\tilde{B}}_{(A)} + \underbrace{\sqrt{2}kK}_{(B)} .$$

The first term $(A)$ upper bounds the difference in loss by the maximum loss $\tilde{B} \times k$ at positions $t$ to $T$ when the sequences do not match. The second term $(B)$ upper bounds the stability differences in predictions in the logit space for positions between $t$ and $T$ that are bounded by $\sqrt{2}kK$ in $\|\cdot\|_2$ norm.

By Definition C.2, we have that the number of distinct tokens in the two suffixes is bounded by $r+1$ for any $r$-rephrasable task. By Theorem E.2, we have the following for any $\epsilon > 0$:

$$\mathbb{P}\left(\left|\mathcal{L}^{(n)}_{\text{in}}(\theta) - \mathcal{L}^{(n)}_{\text{ppt}}(\theta)\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{T\epsilon^2}{2\rho^2}\right) .$$

Therefore, sub-Gaussian concentration holds with $\sigma^2 = \mathcal{O}\left(\frac{\rho}{\sqrt{T}}\right)$.

# I Metric Entropy and $\epsilon$

Yüksel and Flammarion [2025, Section 3.4.] provide metric entropies of general Markov chains of order $k$ and a single-layer self-attention. We recollect their main results below and then comment about metric entropies of other function classes.

In addition, our results in Sections 3 and 4 depend on the optimal choice of $\epsilon$ which is implicitly defined by a fixed point equation, e.g., Equation (5). We expect that $\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T)$ scales with $\log \frac{1}{\epsilon}$ and this choice of $\epsilon$ only adds logarithmic dependencies in the problem parameters. We comment on the choice of $\epsilon$ in the examples below.

**Discrete Markov chains.** For general Markov chains of order $k$ and size $d$, consider the following function class:

$$\mathcal{F} \coloneqq \{f_\theta \coloneqq \theta \circ c : \mathcal{D}^\star \to \mathbb{R}^d \,|\, \theta \in \Theta\} ,$$

where $c(x_1, \ldots, x_T)$ is the function that takes the last $k$ tokens and maps it to a one-hot encoding in $\mathbb{R}^{d^k}$ with 1 at the index corresponding to the last $k$ tokens and 0 otherwise, and $\theta \in \mathbb{R}^{d \times d^k}$ is a matrix. Then, this function class induces the following data generation processes:

$$\mathcal{P} \coloneqq \{p_\theta \coloneqq \sigma \circ f_\theta : \mathcal{D}^\star \to \mathcal{P}(\mathcal{D}) \,|\, \theta \in \Theta\} .$$

Consider the following bounded matrices for the parameter space:

$$\Theta \coloneqq \{\theta \in \mathbb{R}^{d \times d^k} \,|\, \|\theta\|_2 \leq B\} .$$

This models all Markov chains of order $k$ and size $d$ with some minimal probability mass on each state, as determined by the choice of $B$.

Consider the following norm on the parameters:

$$\|\theta\|_{2,\infty} \coloneqq \max_{i=1,\ldots,d^k} \|\theta_i\|_2 .$$

Then, the metric entropy of the function class $\Theta$ is given by:

$$\mathcal{H}_{2,\infty}(\Theta, \epsilon) = \mathcal{O}\left(d^{k+1} \log \frac{B}{\epsilon}\right) .$$

That is, the metric entropy scales linearly with the dimensionality of parameters and logarithmically with the norm of the parameters. In particular, the dependency on the $\epsilon$ is logarithmic. Since the class is Lipschitz with constant $L = 1$, we have that

$$\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) = \mathcal{O}\left(d^{k+1} \log \frac{B}{\epsilon}\right) .$$

The optimal choice of $\epsilon$ then depends at worst linearly on the problem parameters, leading to at worst logarithmic dependencies in the problem parameters.

**Single-layer self-attention.** Let $Q, K, V \in \mathbb{R}^{d \times d}$ be matrices. Let $f_{Q,K,V}$ be the following map:

$$f_{Q,K,V}(\vec{x}) = \sum_{i=1}^{|\vec{x}|} a_{Q,K}(\vec{x}, |\vec{x}|)_i V x_i , \quad \text{where} \quad a_{Q,K}(\vec{x}, t)_i = \begin{cases} \frac{e^{\langle Q x_t, K x_i \rangle}}{\sum_{j=1}^t e^{\langle Q x_t, K x_j \rangle}} & \text{if } i \leq t , \\ 0 & \text{otherwise} . \end{cases} .$$

Then, let the function class be given by:

$$\mathcal{F} \coloneqq \{f_{Q,K,V} : \mathcal{D}^\star \to \mathbb{R}^d \,|\, Q, K, V \in \mathbb{R}^{d \times d}\} .$$

Then, this function class induces the following data generation processes:

$$\mathcal{P} \coloneqq \{p_\theta \coloneqq \sigma \circ f_\theta : \mathcal{D}^\star \to \mathcal{P}(\mathcal{D}) \,|\, \theta \in \Theta\} .$$

Consider the following bounded matrices for the parameter space:

$$\Theta := \{\theta = (Q, K, V) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d} \mid \|Q\|_2 \leq B, \|K\|_2 \leq B, \|V\|_2 \leq B\}\,.$$

This models all single-layer self-attention models with bounded weights.

Consider the following maximum norm on the parameters:

$$\|\theta\|_{2,\infty} := \max(\|Q\|_2, \|K\|_2, \|V\|_2)\,.$$

Then, the metric entropy of the function class $\Theta$ is given by:

$$\mathcal{H}_{2,\infty}(\Theta, \epsilon) = \mathcal{O}\left(d^2 \log \frac{B}{\epsilon}\right)\,.$$

Similarly, the metric entropy scales linearly with the dimensionality of parameters and logarithmically with the norm of the parameters. We can use the fact that softmax is Lipschitz with respect to $\ell_2$ norm to show that $f_{Q,K,V}$ is Lipschitz with constant $L = \mathcal{O}(\sqrt{T} B \sup_{\vec{x}} \|\vec{x}\|_2)$. Therefore, we have that

$$\mathcal{H}_{2,\infty}(\mathcal{F}, \epsilon, T) = \mathcal{O}\left(d^2 \log \frac{B^2\sqrt{T}}{\epsilon}\right)\,.$$

Again, the optimal choice of $\epsilon$ then incurs at worst logarithmic dependencies in the problem parameters.

**Complex function classes.** In order to deal with more complex classes such as transformers, we can use the fact that metric entropy scales additively with the number of layers. That is, when we compose two hypotheses spaces, the metric entropy of the composition is the sum of the metric entropies of the two spaces. In this way, one can incorporate multi-layer perceptrons and other components within transformer blocks and any number of layers.

Overall, we expect the metric entropy to scale linearly with the number of total parameters and logarithmically with the maximum activation norm. Without normalization layers stabilizing the norms of activations, the term based on the input norm can yield a linear dependency on the number of layers. This is due to the fact that the norm can grow multiplicatively with the number of layers.

## J Further Discussion on Assumptions

**Mixing and Rephrasability.** Mixing is a classical concept in problems with weak dependence [Doukhan, 1995]. Behavior of the mixing coefficients are well-studied for Markov chains and some non-Markovian processes [Bradley, 2005]. Therefore, we focus on the rephrasability condition.

First, we explain how the rephrasability condition in Definition C.2 subsumes the mixing condition in Definition 3.8. In particular, we have the following inequality:

$$\sup_{\vec{x},\vec{y}\in\tilde{\mathcal{D}}^k} \inf_{\gamma\in\Gamma_{\vec{z}}(\vec{x},\vec{y})} \mathbb{E}_{\vec{\mu},\vec{\nu}\sim\gamma}\left[d_H(\vec{\mu},\vec{\nu})\right] \leq \min_{\epsilon\in(0,1)} \phi_{\vec{z}}(\epsilon) + T\epsilon\,.$$

This is due to the fact that the coupling $\gamma$ be chosen to verify these two conditions: (i) coalesced sequences with weight at least $1-\epsilon$ after $\phi(\epsilon)$ steps evolve together, and (ii) the $\epsilon$ fraction that do not coalesce evolve independently which incurs at most $T\epsilon$ in the distance. Note that the term $\phi(\epsilon) + T\epsilon$ is the same as the term $\rho$ in Theorem 3.9 for the mixing condition except the stability term. Therefore, if the mixing condition is satisfied with a good constant, so is the rephrasability condition. In this sense, rephrasability is a generalization of the mixing condition. However, the reverse is not true as the rephrasability condition is not local.

**Non-mixing Examples.** Many algorithmic tasks are non-mixing as the output, which is at the end of the sequence, is a deterministic function of the whole input. Consider the task of parity computation given strings composed of zeros and ones. The data generation process at the start is random, e.g., each digit is sampled with a Bernouilli distribution. The very final token that is the output of the task is a function of the whole sequence. Therefore, there is no mixing as any change in digits results in a change at the end of the sequence. However, the rephrasability condition is verified with $r = 1$. Similarly, consider the task of addition. Again, there is no mixing but the rephrasability condition is verified with $r$ that captures the maximum number of digit carries within the dataset. We expect language to have a similar behavior as it has long-range dependencies, e.g., a character in a book. Therefore, bounds based on rephrasability goes beyond the mixing results in terms of applicability.

**Estimation of rephrasability.** We discuss how to get empirical estimates of the rephrasability condition when one has access to the ground-truth process or to a sufficently accurate approximation thereof for sampling trajectories.g An interesting example is a large-scale language model that serves as a proxy for natural language. In such cases, we can estimate the rephrasability condition by solving a discrete optimal transport problem over sampled trajectories as follows. Let $H, M$ be fixed numbers.

1. Sample the prompt $\vec{z}$.
2. Sample two continuations $\vec{x}, \vec{y} \in \tilde{\mathcal{D}}^k$ of the prompt $\vec{z}$.
3. Sample $H$ completions of $\vec{z} \circ \vec{x}$ to the full length $T$, denoted by $\vec{x}_1, \dots, \vec{x}_H$.
4. Sample $H$ completions of $\vec{z} \circ \vec{y}$ to the full length $T$, denoted by $\vec{y}_1, \dots, \vec{y}_H$.
5. Solve the optimal transport problem of carrying the empirical distribution $\{\vec{x}_1, \dots, \vec{x}_H\}$ to the empirical distribution $\{\vec{y}_1, \dots, \vec{y}_H\}$ with the Hamming distance over the sequences, $d(\vec{a}, \vec{b}) = \sum_{i=1}^{T} 1_{\vec{a}_i \neq \vec{b}_i}$.
6. Repeat the above procedure $M$ times.

The optimal transport problem can be solved with Sinkhorn-Knopp algorithm. The range of the costs obtained gives us the plausible range for $r$. While the tractability of this approach in realistic settings remains unclear, we view this as an exciting and promising research direction that would require a dedicated empirical investigation.

# K Concentration of training mixture

In this section, we explain how fast $p^\star_N$ approximates $p^\star_\infty$. In Section 4, we discussed the how the following sampling procedure

$$w \sim \mathcal{P}_\mathcal{W}\,, \quad \vec{z} \sim \pi^\star_w\,, \quad \forall t \in [T]: \quad x_t \sim p^\star_w(\cdot \mid \vec{x}_t)\,,$$

is equivalent to the following:

$$\vec{z} \sim \pi^\star_\infty\,, \quad \forall t \in [T]: \quad x_t \sim p^\star_\infty(x \mid \vec{x}_t)\,.$$

More generally, for any prior distribution over the tasks $\hat{\mathcal{P}}_\mathcal{W}$, the following sampling procedure

$$w \sim \hat{\mathcal{P}}_\mathcal{W}\,, \quad \vec{z} \sim \pi^\star_w\,, \quad \forall t \in [T]: \quad x_t \sim p^\star_w(\cdot \mid \vec{x}_t)\,,$$

is equivalent to the following:

$$\vec{z} \sim \hat{\pi}\,, \quad \forall t \in [T]: \quad x_t \sim \hat{p}(x \mid \vec{x}_t)$$

where $\hat{\pi}$ and $\hat{p}$ are defined as follows:

$$\hat{\pi}(\vec{z}) = \int_{w \in \mathcal{W}} \pi^\star_w(\vec{z}) \hat{\mathcal{P}}_\mathcal{W}(w) dw\,,$$
$$\hat{p}(x \mid \vec{x}_t) = \int_{w \in \mathcal{W}} p^\star_w(x \mid \vec{x}_t)\, \hat{\mathcal{P}}_\mathcal{W}(w \mid \vec{x}_t)\, dw\,,$$

where $\hat{\mathcal{P}}_\mathcal{W}(w \mid \vec{x}_t)$ is the posterior distribution over tasks given the sequence $\vec{x}_t$:

$$\hat{\mathcal{P}}_\mathcal{W}(w \mid \vec{x}_t) \propto \hat{\mathcal{P}}_\mathcal{W}(w) \mathcal{P}_w(\vec{x}_t)\,, \quad \text{with} \quad \mathcal{P}_w(\vec{x}_t) \coloneqq \pi^\star_w(\vec{z}) \prod_{i=1}^{t-1} p^\star_w(x_i \mid \vec{x}_i)\,.$$

It is easy to see this equivalence by an induction argument. The marginal distribution over $\vec{z}$ clearly matches. Then, for any $t \geq 1$, we have

$$\begin{aligned}\hat{p}(x_t \circ \vec{x}_t) &= \hat{p}(x_t \mid \vec{x}_t)\hat{p}(\vec{x}_t) \\ &= \int_{w\in\mathcal{W}} p^\star_w(x \mid \vec{x}_t)\hat{\mathcal{P}}_{\mathcal{W}}(w \mid \vec{x}_t)\hat{p}(\vec{x}_t)dw \\ &= \int_{w\in\mathcal{W}} p^\star_w(x \mid \vec{x}_t)\mathcal{P}_w(\vec{x}_t)\hat{\mathcal{P}}_{\mathcal{W}}(w)dw \\ &= \int_{w\in\mathcal{W}} p^\star_w(x \circ \vec{x}_t)\hat{\mathcal{P}}_{\mathcal{W}}(w)dw\,.\end{aligned}$$

The loss incurred by $\hat{p}$ is given by:

$$\mathcal{L}_{\text{out}}(\hat{p}) = \mathbb{E}_{\vec{x}_{T+1}}\left[-\log\hat{p}(\vec{x}_{T+1}) + \log p^\star_\infty(\vec{x}_{T+1})\right] = \mathrm{KL}(p^\star_\infty \parallel \hat{p})\,.$$

Let $q^\star(w, \vec{x}_{T+1}) = \mathcal{P}_{\mathcal{W}}(w)\mathcal{P}_w(\vec{x}_{T+1})$ and $\hat{q}(w, \vec{x}_{T+1}) = \hat{\mathcal{P}}_{\mathcal{W}}(w)\mathcal{P}_w(\vec{x}_{T+1})$. By the chain rule of KL divergence,

$$\mathrm{KL}(q^\star \parallel \hat{q}) = \mathrm{KL}(p^\star_\infty \parallel \hat{p}) + \mathbb{E}_{\vec{x}_{T+1}}\left[\mathrm{KL}(\mathcal{P}_{\mathcal{W}}(w \mid \vec{x}_{T+1}) \parallel \hat{\mathcal{P}}_{\mathcal{W}}(w \mid \vec{x}_{T+1}))\right]\,.$$

Note that the left-hand side is equal to KL divergence of the prior over the tasks:

$$\begin{aligned}\mathrm{KL}(q^\star \parallel \hat{q}) &= \int_{w\in\mathcal{W}}\int_{\vec{x}_{T+1}} \mathcal{P}_{\mathcal{W}}(w)\mathcal{P}_w(\vec{x}_{T+1})\log\frac{\mathcal{P}_{\mathcal{W}}(w)\mathcal{P}_w(\vec{x}_{T+1})}{\hat{\mathcal{P}}_{\mathcal{W}}(w)\mathcal{P}_w(\vec{x}_{T+1})}d\vec{x}_{T+1}dw \\ &= \int_{w\in\mathcal{W}} \mathcal{P}_{\mathcal{W}}(w)\log\frac{\mathcal{P}_{\mathcal{W}}(w)}{\hat{\mathcal{P}}_{\mathcal{W}}(w)}dw \\ &= \mathrm{KL}(\mathcal{P}_{\mathcal{W}} \parallel \hat{\mathcal{P}}_{\mathcal{W}})\,.\end{aligned}$$

Since KL is non-negative,

$$\mathrm{KL}(q^\star \parallel \hat{q}) = \mathrm{KL}(\mathcal{P}_{\mathcal{W}} \parallel \hat{\mathcal{P}}_{\mathcal{W}}) \geq \mathrm{KL}(p^\star_\infty \parallel \hat{p}) = \mathcal{L}_{\text{out}}(\hat{p})\,.$$

Therefore, the loss of $\hat{p}$ is controllable by the KL divergence of the prior over the tasks.

In order to turn the computations above into a result on $p^\star_N$, we need to smooth it so that it is supported over all tasks. This is possible under a Lipschitzness assumption on $\mathcal{P}_w(\vec{x}_{T+1})$ in the parameter space $w \in \mathcal{W}$. That is, there must exist a constant $L$ such that

$$\mathcal{L}_{\text{out}}(p^\star_N) \leq \mathcal{L}_{\text{out}}(\hat{p}_N) + L\eta\,,$$

where $\hat{p}_N$ is some smoothed version of $p^\star_N$ with parameter $\eta$ and kernel $K$, e.g.,

$$\hat{p}_N(\cdot) = \int_{w\in\mathcal{W}} p^\star_w(\cdot)\hat{\mathcal{P}}_{\mathcal{W}}(w)dw\,, \quad \text{where} \quad \hat{\mathcal{P}}_{\mathcal{W}}(w) = \frac{1}{N\eta}\sum_{i=1}^{N} K\left(\frac{w - w_i}{\eta}\right)\,.$$

Then, the speed in which $p^\star_N$ approximates $p^\star_\infty$ is simply the classical question of kernel density estimation within the task space $\mathcal{W}$. In particular, we have that

$$\mathcal{L}_{\text{out}}(p^\star_N) \leq \mathcal{L}_{\text{out}}(\hat{p}_N) + L\eta \leq \mathrm{KL}(p^\star_\infty \parallel \hat{p}_N) + L\eta\,.$$