THE COVERAGE PRINCIPLE: HOW PRE-TRAINING ENABLES POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models demonstrate remarkable abilities when pre-trained on large text corpora and fine-tuned for specific tasks, but how and why pre-training shapes the success of the final model remains poorly understood. Notably, although pre-training success is often quantified by cross entropy loss, cross entropy can be poorly predictive of downstream performance. Instead, we provide a theoretical perspective on this relationship through the lens of *coverage*, which quantifies the probability mass the pre-trained model places on high-quality responses and which is necessary and sufficient for post-training and test-time scaling methods like Best-of-N to succeed. Our main results develop an understanding of *the coverage principle*, a phenomenon whereby next-token prediction implicitly optimizes toward a model with good coverage. In particular, we uncover a mechanism that explains the power of coverage in predicting downstream performance: *coverage generalizes faster than cross entropy*, avoiding spurious dependence on problem dependent parameters such as the sequence length. We also study practical algorithmic interventions with provable benefits for improving coverage, including (i) model/checkpoint selection procedures, (ii) gradient normalization schemes, and (iii) test-time decoding strategies.

1 INTRODUCTION

The remarkable capabilities of language models stem from a two-stage training process: (1) large-scale pre-training via next-token prediction with the cross-entropy loss (predicting what token should follow a prefix) and (2) targeted post-training—typically via reinforcement learning—to adapt the model to specific domains and tasks. Investing more compute and data into pre-training often enables post-training to produce a stronger model, but theoretical understanding of how these stages interact is limited. Indeed, despite substantial investment into scaling pre-training (Gadre et al., 2025; Sardana et al., 2024; Hoffmann et al., 2022), several works have demonstrated that starting post-training from a better next-token predictor does not ensure stronger performance on downstream tasks (Liu et al., 2022; Zeng et al., 2025; Chen et al., 2025; Lourie et al., 2025). Motivated by this disconnect, we theoretically investigate the connection between pre-training objectives and downstream success, asking:

Can we precisely characterize the relationship between the next-token prediction loss and downstream performance? What metrics are most predictive of downstream success?

Motivated by the recent interest in test-time scaling, we focus our attention on post-training via Best-of-N (BoN) sampling or reinforcement learning with verifiable rewards. For a prompt x, Best-of-N draws n responses y from the model and returns the best response according to a task-specific reward. Several prior works have demonstrated that the performance of BoN is strongly indicative of how well the model will perform after post-training via reinforcement learning (Yue et al., 2025; Wu et al., 2025).

Our starting point is the observation that cross-entropy alone cannot provide meaningful answers to the questions above; see Figure 1, which illustrates that cross-entropy can be *anti-correlated* with BoN performance, echoing Chen et al. (2025). Instead, we show that the missing link is the *coverage profile*, a novel refinement of cross-entropy that explicitly quantifies the model's ability to assign sufficient probability mass to rare but high-quality responses.

Definition 1.1 (Coverage profile)**.** *The coverage profile of a model $\widehat{\pi}$ for a distribution π is*

$$\mathtt{Cov}_N(\pi \,\|\, \widehat{\pi}) := \mathbb{P}_{x\sim\mu, y\sim\pi(\cdot|x)}\left[\frac{\pi(y \mid x)}{\widehat{\pi}(y \mid x)} \geq N\right], \tag{1}$$

where $N \geq 1$ is the number of Best-of-N sampling attempts.

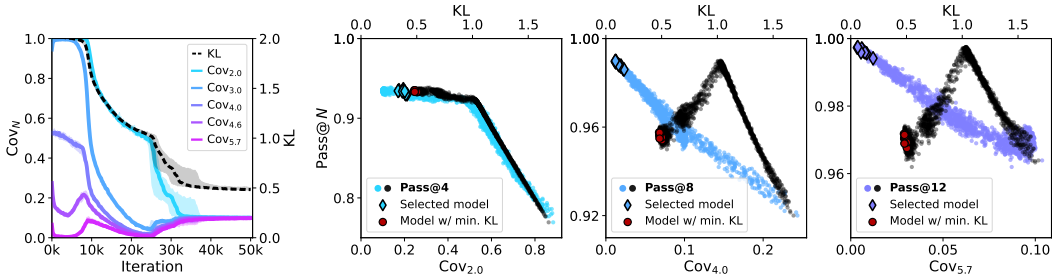

Figure 1: **The coverage profile predicts** Pass@N **better than KL divergence.** We train models in a graph reasoning task and record KL divergence, coverage profile (both measured w.r.t., π_{D}), and Pass@N performance; see Appendix G for details. Left: Convergence of coverage and KL divergence over training, showing that KL improves monotonically but coverage can *degrade* with training. Right: Scatter plots of KL (top axis), $\mathsf{Cov}_{N/2}$ (lower axis) and Pass@N of checkpoints. Although KL and Cov_N exhibit comparable predictive power for small N, Cov_N is a better predictor for large N. Also visualized are checkpoints selected via the tournament procedure of Eq. (13) (marked $\diamondsuit$) and by minimizing KL (marked red), demonstrating that the former selects better models for Pass@N.

Here, y is the full response when prompted with x, π represents the pre-training data distribution, which we presuppose covers downstream tasks of interest, and $\widehat{\pi}$ is the pre-trained model. We prove that a **good coverage profile is necessary and sufficient for Best-of-N to succeed** (see Section 2, as well as Propositions D.6 and D.7). This is highlighted in Figure 1, where we find that the coverage profile is correlated with downstream performance for Best-of-N (which is exactly Pass@N), even when cross-entropy is not.[1] Motivated by this characterization of BoN performance, we ask: *When, and through what mechanism, does next-token prediction produce a model $\widehat{\pi}$ with good coverage?*

1.1 Contributions

We develop a theoretical understanding of **the coverage principle**, whereby next-token prediction implicitly optimizes toward a model with a good coverage profile, inheriting the training corpus' coverage over tasks of interest.

Cross-entropy: Scaling laws and limitations (Section 3). We begin by deriving provable scaling laws that link cross-entropy—specifically, a certain sequence-level notion—to coverage and hence downstream performance, but show that cross-entropy can be sensitive to sequence length and other problem parameters, leading to vacuous predictions; this motivates our main results.

Next-token prediction implicitly optimizes coverage (Section 4). The first of our main theoretical results (Theorem 4.1) is a new generalization analysis for next-token prediction (more generally, maximum likelihood) that exploits the unique structure of the logarithmic loss to show that **coverage can generalize faster than cross entropy**; we refer to this as the coverage principle. Concretely, our analysis shows that the coverage profile for models learned with next-token prediction (i) avoids spurious dependence on problem-dependent parameters such as sequence length (in contrast to cross-entropy), and (ii) converges *faster* still as the tail parameter N is increased. Our analysis—which is similar in spirit to Mendelson's *small ball method* (Mendelson, 2014; 2017)—can be viewed as a giving a new, fine-grained understanding of maximum likelihood.

Stochastic gradient descent through the lens of coverage (Section 5). The preceding results apply to general model classes Π, but consider the empirical maximizer of the next-token prediction (maximum likelihood) objective, in the vein of classical techniques in learning theory. For the second of our main results, we focus on a specific model class—overparameterized autoregressive linear models (2)—but take a more realistic approach and analyze stochastic gradient descent (SGD) on the next-token prediction objective, in the one-pass ("compute-optimal") regime. We show that while SGD provably optimizes the coverage profile, it experiences suboptimal dependence on the sequence length H. We then show that *gradient normalization* (which is loosely connected to Adam-like updates (Bernstein & Newhouse, 2024)) provably improves coverage, removing dependence on the sequence length.

[1] Formally, the coverage profile refines cross-entropy/KL divergence; the former is the cumulative distribution function (CDF) of the log density ratio $\log \frac{\pi(y|x)}{\widehat{\pi}(y|x)}$, while KL divergence is the mean; see Remark C.1.

Interventions for better coverage (Section 6). Finally, we look beyond standard next-token prediction and explore families of new interventions aimed at improving coverage in theory.
(i) Test-time (Section 6.1). We show that for standard token-level SGD, a new test-time decoding strategy inspired by *test-time training* (Sun et al., 2020; Akyürek et al., 2025) provably improves coverage.
(ii) Model/checkpoint selection (Section 6.2). For selecting the best model (or checkpoint) from a small number of candidates, we give *tournament* procedures that enjoy significantly better coverage profile (particularly with respect to the tail parameter N) than naïve validation with cross-entropy.

Additional results (Appendix F). Beyond the results above, we show that: (1) MLE can find models with low coverage even in the presence of severe misspecification; (2) coverage can generalize better under additional structural properties of the model class such as convexity (Appendix F.2).

In summary, we believe that coverage offers a new perspective on the connection between pre-training objectives and downstream post-training success. Our results demonstrate that this perspective is mathematically rich and fundamental, opening the door to a deeper understanding; cf. Appendix A.

2 Problem Setup

We now introduce the formal problem setup for the remainder of the paper.

Next-token prediction and maximum likelihood. We work in the following setting, which subsumes next-token prediction: $\mathcal{X}$ is the prompt space, $\mathcal{Y}$ is the response space, and $\pi_{\mathsf{D}} : \mathcal{X} \to \Delta(\mathcal{Y})$ is the data distribution. We are given a dataset $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^n$ where $x^i \sim \mu$ and $y^i \sim \pi_{\mathsf{D}}(\cdot \mid x^i)$. We consider the maximum likelihood objective $\widehat{L}_n(\pi) := \sum_{i=1}^n \log \pi(y^i \mid x^i)$, and refer to $\widehat{\pi} := \arg\max_{\pi \in \Pi} \widehat{L}_n(\pi)$ as the *maximum likelihood estimator* for a user-specified model class Π. This is a generalization of the next-token prediction, where $\mathcal{Y} = \mathcal{V}^H$ is a token sequence and $\pi(y \mid x) = \prod_{h=1}^H \pi(y_h \mid x, y_{1:h-1})$ is explicitly autoregressive, so that $\widehat{L}_n(\pi) = \sum_{i=1}^n \sum_{h=1}^H \log \pi(y^i_h \mid x^i, y^i_{1:h-1})$. We specialize to next-token prediction at certain points but otherwise focus on the general setting. We make the following realizability assumption throughout.

Assumption 2.1 (Realizability)**.** *The data distribution π_{D} is realizable by some model $\pi \in \Pi$.*

This formulation captures pre-training and SFT, with some caveats; see Appendix A.1.

Post-training and the coverage profile. Given a reward function $r_{\mathsf{T}}(x, y) \in \{0, 1\}$ representing success at a downstream task T, the goal is to fine-tune $\widehat{\pi}$—through reinforcement learning or test-time scaling—to obtain near-optimal reward. We show (Propositions D.6 and D.7) that for any task-specific comparator policy $\pi_{\mathsf{T}} : \mathcal{X} \to \Delta(\mathcal{Y})$, Best-of-N sampling with $\widetilde{\Theta}(N)$ samples satisfies $\mathbb{E}_{x\sim\mu}[r_{\mathsf{T}}(x, \pi_{\mathsf{T}}(x)) - r_{\mathsf{T}}(x, \widehat{\pi}^{\mathsf{BoN}}_N(x))] \asymp \mathsf{Cov}_N(\pi_{\mathsf{T}} \,\|\, \widehat{\pi})$, so a good coverage profile for π_{T} is sufficient for high reward. Moreover, in a worst-case sense a good coverage profile is necessary for high reward; see Proposition D.7. Further, while less well understood, some form of coverage is thought to be necessary for the success of reinforcement learning methods like GRPO (Yue et al., 2025; Song et al., 2024).

Returning to pre-training, it is clear that there is little hope that next-token prediction will produce a model $\widehat{\pi}$ with good coverage with respect to a downstream task unless the data distribution π_{D} itself has reasonable coverage with respect to this task. We therefore posit that the data distribution covers such a downstream task, in the sense that it includes high-reward responses with some bounded-below probability. Since coverage satisfies a transitivity property, it follows that coverage with respect to π_{D} implies coverage with respect to the optimal policy for the downstream task. For example, if π_{D} has a 10% chance of generating a correct response, and $\mathsf{Cov}_{N/10}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) = \varepsilon$, then we get 10ε error.[2] Thus, **going forward, we focus on understanding when next-token prediction achieves good coverage $\mathsf{Cov}_N(\pi_{\mathsf{D}} \,\|\, \widehat{\pi})$ relative to the data distribution π_{D} itself,** and avoid concerning ourselves with specific details of the task policy π_{T} or the specific relationship between π_{T} and π_{D}.

Autoregressive linear models. We analyze next-token prediction and maximum likelihood for general model classes Π, but our running example throughout the paper will be the class Π of *autoregressive linear models*, defined by a known feature map $\phi : \mathcal{X} \times \mathcal{V}^\star \to \mathbb{R}^d$. For each parameter $\theta \in \Theta \subset \mathbb{R}^d$, the model $\pi_\theta = (\pi_\theta)_{h=1}^H$ is defined by

$$\pi_\theta(y_h \mid x, y_{1:h-1}) \propto \exp(\langle \theta, \phi(x, y_{1:h}) \rangle). \tag{2}$$

[2] See Proposition D.5 for formal results.

In practice, autoregressive sequence models—such as those based on transformers—generate each token by sampling from a softmax distribution whose logits are given by a linear combination of learned features (Radford et al., 2019). Eq. (2) simplifies this by freezing the feature map, yet remains expressive enough to model complex non-Markovian dependencies, depending on the choice of features.

Assumption 2.2. *We assume that* $\Theta \subseteq \{\theta : \|\theta\| \leq 1\}$ *is convex, and* $\sup_{h,x,y_{1:h}} \|\phi(x, y_{1:h})\| \leq B$.

3 Cross-Entropy and Coverage: Scaling Laws and Limitations

A natural approach to understanding when next-token prediction achieves good coverage is to appeal to cross-entropy—perhaps first showing that next-token prediction achieves low cross-entropy (which is true asymptotically), and then relating cross-entropy to coverage. In this section we motivate our main results by showing that while this is possible in a weak sense, it does not yield predictive guarantees for downstream performance in the finite-sample regime.

Define the *sequence-level* cross-entropy for $\widehat{\pi}$ as $D_{\mathsf{CE}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) := \mathbb{E}_{\pi_{\mathsf{D}}}\left[\sum_{h=1}^{H} \log \frac{1}{\widehat{\pi}(y_h|x,y_{1:h-1})}\right]$. Since $\mathbb{E}_{\mathcal{D}\overset{\text{i.i.d.}}{\sim}\pi_{\mathsf{D}}}\left[\widehat{L}_n(\pi)\right] = -n \cdot D_{\mathsf{CE}}(\pi_{\mathsf{D}} \,\|\, \pi)$, one expects that as we scale up compute, number of samples n, and model capacity Π, $D_{\mathsf{CE}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) \to D_{\mathsf{CE}}(\pi_{\mathsf{D}} \,\|\, \pi_{\mathsf{D}})$, or equivalently $D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) \to 0$, where $D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) := \mathbb{E}_{\pi_{\mathsf{D}}}\left[\sum_{h=1}^{H} \log \frac{\pi_{\mathsf{D}}(y_h|x,y_{1:h-1})}{\widehat{\pi}(y_h|x,y_{1:h-1})}\right]$ is the sequence-level KL divergence.

A simple scaling law for cross-entropy. We show below that if that the model $\widehat{\pi}$ has reasonable KL divergence to the data distribution, the coverage profile can be bounded:

Proposition 3.1 (KL-to-coverage; see Proposition D.1)**.** *For all* $N \geq e$, $\mathsf{Cov}_N(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) \leq \frac{D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi})}{\log(N/e)}$.

Combining Proposition 3.1 with Proposition D.6 and our assumption that π_{D} has good coverage with respect to the downstream task yields a simple "scaling law" for test-time compute with BoN:

> Consider a task of interest with reward $r_{\mathsf{T}}(x, y)$, and suppose the data distribution π_{D} itself has constant probability of success (i.e., sampling $y \sim \pi_{\mathsf{D}}(\cdot \mid x)$ with $r_{\mathsf{T}}(x, y) = 1$). To achieve sub-optimality ε with Best-of-N, it suffices to choose the compute budget N as
>
> $$N \approx \exp\left(\frac{D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi})}{\varepsilon}\right). \tag{3}$$

That is, for a fixed model $\widehat{\pi}$ and KL-divergence level $D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) \leq D_{\mathsf{CE}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi})$, Eq. (3) predicts that test-time compute should increase exponentially with the desired accuracy ε.[3]

Insufficiency of cross-entropy. At first glance, this seems to be in line with empirical test-time scaling laws (OpenAI, 2024), but there is an issue: While *token-level* cross-entropy has been observed to be modest in contemporary language models (Kaplan et al., 2020; Hoffmann et al., 2022; Xia et al., 2022), the *sequence-level* cross-entropy (and KL-divergence) generally grows with the length H of the sequence, so that Eq. (3) predicts exponential test-time scaling in the sequence length. Moreover, such a law cannot hold if we only assume token-level cross-entropy is bounded; see Proposition D.7.

Is this the end of the story? On the one hand, it is simple to show (Proposition D.2) that Proposition 3.1 is tight for a worst-case pair of models. Moreover, even for the autoregressive linear model in Eq. (2), sequence-level KL divergence scales linearly with the sequence length H, as shown in the next result.

Proposition 3.2. *Fix* $H \in \mathbb{N}$. *There exists* $\phi : \mathcal{X} \times \mathcal{V}^\star \to \mathbb{B}_2(1)$ *and induced autoregressive linear class* Π *with parameter space* $\Theta = \mathbb{B}_2(1)$*, distribution* μ *over* $\mathcal{X}$ *and data distribution* $\pi_{\mathsf{D}} \in \Pi$*, such that for any* proper *estimator* $\widehat{\pi} = \widehat{\pi}(\mathcal{D}) \in \Pi$*, it holds that w.p. at least* 0.5*,* $D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) \geq \Omega(\frac{H}{n})$.

This behavior is reflected empirically in Figure 2 in a graph reasoning task. Yet, for this task (Figure 2), we find that, in spite of large cross-entropy/KL, next-token prediction learns a model $\widehat{\pi}$ with a good coverage profile across a range of sequence lengths and that downstream Best-of-N succeeds. Why is this happening? In light of the discussion above, it must be related to specific inductive bias of the next-token prediction objective itself.

[3]Neither KL divergence nor the coverage profile are observable quantities (though cross entropy is an estimable upper bound on KL), so this is a theoretical prediction rather than a practical one as-is; see Remark C.2.

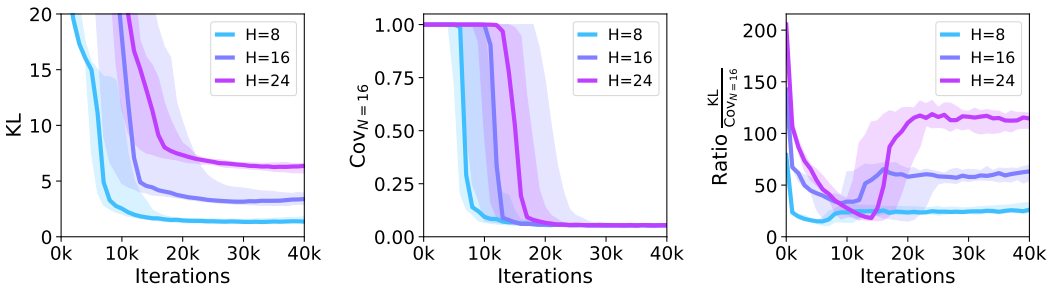

Figure 2: **The coverage profile avoids spurious dependence on sequence length.** We train models in a graph reasoning task and record their KL divergence and coverage profile, measured w.r.t., π_{D} as we vary the problem horizon (sequence length); see Appendix G for details. Left: Convergence of KL over training for three horizons H, demonstrating that KL at convergence scales linearly in the horizon H. Center: Convergence of Cov_N over training, manifesting no dependence on H at convergence. Right: Ratio of KL over Cov_N, showing that Proposition 3.1 can be overly conservative.

A glimmer of hope: Case study in Bernoulli models. To see why large cross-entropy may not be a barrier to coverage, consider perhaps the simplest setting, *Bernoulli models*, where $\mathcal{X} = \{\perp\}$, $\mathcal{Y} = \{0, 1\}$, $\Pi = \{\mathrm{Ber}(p)\}_{p\in(0,1/2)}$, and $\pi_{\mathsf{D}} = \mathrm{Ber}(p^\star)$ for some small $p^\star \in (0, 1/2)$.

The maximum likelihood model is $\widehat{\pi} = \mathrm{Ber}(\widehat{p})$, where $\widehat{p}$ is the empirical frequency of $y = 1$ in the dataset. We observe that with positive probability (and constant probability if $n \leq 1/p^\star$), the dataset $\mathcal{D}$ will only contain examples where $y = 0$, so that the maximum likelihood model is $\widehat{\pi} = \mathrm{Ber}(0)$. This implies that expected KL divergence is infinite: $\mathbb{E}[D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi})] = +\infty$. However, the coverage profile turns out to be well-behaved; a direct calculation shows that $\mathsf{Cov}_N(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) \lesssim \frac{\log(\delta^{-1})}{n}$ with probability at least $1-\delta$ for all $N \geq 2$; this gives hope that even though cross-entropy itself is infinite, maximum likelihood may actually learn a model with good coverage in the background. In what follows, we will show that this is not a fluke, but a general phenomenon.

Remark 3.1 (Missing mass)**.** *The underlying issue is one of* missing mass*: there are responses that even a well-generalizing learner will fail to cover, and for these we may incur a large contribution to the KL-divergence. More generally, KL-divergence and cross-entropy are susceptible to contributions of the scale* $\log W_{\mathsf{max}}$ *where* $W_{\mathsf{max}} = \max_{\pi\in\Pi} \left\| \frac{\pi_{\mathsf{D}}}{\pi} \right\|_\infty$ *(which could be as large as* H*, as in Proposition 3.2) when the model does not have enough information to generalize/extrapolate. This phenomenon is particularly pronounced when the prompt distribution is heterogeneous.*

4 Next-Token Prediction Implicitly Optimizes Coverage

We now present our main result, which establishes the *coverage principle*: due to the unique structure of the logarithmic loss, maximum likelihood can learn models with good coverage even when cross-entropy is vacuously large. We make use of the following covering number.

Definition 4.1. *For a class* Π *and* $\alpha \geq 0$*, we let* $\mathcal{N}_\infty(\Pi, \alpha)$ *denote the size of the smallest cover* Π' *such that for all* $\pi \in \Pi$*, there exists* $\pi' \in \Pi'$ *such that* $\sup_{x\in\mathcal{X}, y\in\mathcal{Y}} |\log \pi(y \mid x) - \log \pi'(y \mid x)| \leq \alpha$.

Theorem 4.1 (Coverage principle)**.** *Let* $N \geq 8$ *be given and* $c > 0$ *be an absolute constant. Suppose Assumption 2.1 holds. With probability at least* $1-\delta$*, the maximum likelihood estimator satisfies*

$$\mathsf{Cov}_N(\widehat{\pi}) \lesssim \frac{1}{\log N} \cdot \underbrace{\inf_{\varepsilon>0}\left\{ \frac{\log \mathcal{N}_\infty(\Pi, \varepsilon)}{n} + \varepsilon \right\}}_{=: \mathcal{C}_{\mathsf{fine}}(\Pi, n)} + \underbrace{\frac{\log \mathcal{N}_\infty\left(\Pi; c \log N\right) + \log(\delta^{-1})}{n}}_{=: \mathcal{C}_{\mathsf{coarse}}(\Pi, N, n)}. \tag{4}$$

Theorem 4.1 has two terms: a *coarse-grained term* $\mathcal{C}_{\mathsf{coarse}}(\Pi, N, n)$ and *fine-grained term* $\mathcal{C}_{\mathsf{fine}}(\Pi, n)$; we interpret each term below.

Fine-grained term. $\mathcal{C}_{\mathsf{fine}}(\Pi, n)$ evaluates the covering number $\mathcal{N}_\infty(\Pi, \varepsilon)$ at a small scale ε (typically $\varepsilon \approx \mathrm{poly}(1/n)$), which matches typical bounds for conditional density estimation (e.g., Bilodeau et al. (2023)) in KL divergence; however, unlike KL-based bounds this term has *no explicit dependence on sequence length* H *or density ratios* $\log W_{\mathsf{max}}$. The term is further scaled by $1/\log N$, which implies that *coverage enjoys faster convergence as we move further into the tail* by increasing N; this reflects the unique structure of the logarithmic loss, and may be viewed as a new form of implicit bias.

Summarizing, the fine-grained term witnesses the *coverage principle*: coverage enjoys faster generalization than cross-entropy; roughly, the rate is what we would expect (via Proposition 3.1) if we could somehow control KL without paying for the sequence length H or density ratio $\log W_{\max}$. See Appendix E for a detailed comparison to standard (asymptotic and non-asymptotic) generalization bounds for maximum likelihood based on Hellinger distance and KL-divergence.

Coarse-grained term. The coarse-grained term $\mathcal{C}_{\texttt{coarse}}(\Pi, N, n)$ captures the *missing mass* phenomenon exemplified by the Bernoulli example in the prequel. This term is not explictly normalized by $1/\log N$ (compared to the fine-grained term), but depends on the covering number $\mathcal{N}_\infty(\Pi, \alpha)$ only at a very large scale $\alpha \approx \log N$. This implies that the dependence on the capacity of Π in this term vanishes as we increase N.

Overall, while the guarantee in Eq. (4) might look surprising at first glance (particularly the coarse term, as we are not aware of any existing generalization bounds with dependence on covering numbers at such a large scale), we show in Proposition F.1 (Appendix J) that both terms are tight in general.

Overview of analysis. The proof of Theorem 4.1 is given in Appendix J (with a high-level sketch in Appendix J.1). The basic idea is to interpret the condition $\texttt{Cov}_N(\pi) \geq \varepsilon$ as a small ball-like *anti-concentration* condition in the vein of Mendelson (2014; 2017). That is, for models π where coverage is large, the condition $\texttt{Cov}_N(\pi) \geq \varepsilon$ witnesses a *one-sided* bound which implies that the empirical likelihood of π is *not too large* with high probability, and thus π cannot be a maximum-likelihood solution.

The coarse-grained term $\mathcal{C}_{\texttt{coarse}}(\Pi, N, n)$ enters because we only need to show that the coverage profile concentrates, not the log loss itself. The fine-grained term $\mathcal{C}_{\texttt{fine}}(\Pi, n)$ enters from one-sided concentration of the empirical likelihood, with the $1/\log N$ scaling arising from the following form of implicit bias: If an example (x^i, y^i) is such that $\pi_{\texttt{D}}(y^i|x^i)/\pi(y^i|x^i) \geq N$, this witnesses a negative contribution of order $\log N$ to the difference $\widehat{L}_n(\pi) - \widehat{L}_n(\pi_{\texttt{D}})$.

Discussion. We emphasize that while covering numbers are a fundamental and widely used noton of capacity in statistical learning and estimation (van de Geer, 2000; Zhang, 2002; Rakhlin & Sridharan, 2012; Bilodeau et al., 2023), they are conservative from a modern generalization perspective. Nonetheless, Theorem 4.1 shows that they are sufficient to capture rich aspects of generalization for coverage, and we expect that our core analysis techniques can be combined with contemporary advances in generalization theory for overparameterized models (Belkin et al., 2019; Bartlett et al., 2020).

4.1 Examples

To build intuition, we analyze the behavior of Theorem 4.1 under a growth assumption on the covering number, then discuss how autoregressive linear models exemplify the coverage principle.

Corollary 4.1. ***(i) Parametric regime:*** *Suppose that there are parameters $d \geq 2$ and $C \geq 2$ such that $\log \mathcal{N}_\infty(\Pi, \alpha) \leq d \log(C/\alpha)$ for $\alpha \in (0, C/2]$. Then for any $N \geq 8$, with probability at least $1-\delta$, $\texttt{Cov}_N(\widehat{\pi}) \lesssim \frac{d\left[\left[\log(C/\log N)\right]_+ + \frac{\log(Cn)}{\log N}\right] + \log(1/\delta)}{n}$.*

(ii) Nonparametric regime: *Suppose that there are parameters $C \geq 2$ and $p > 0$ such that $\log \mathcal{N}_\infty(\Pi, \alpha) \leq (C/\alpha)^p$ for $\alpha \in (0, C/2]$. Then for any $N \geq 8$ and $n \geq \log^{1/p} N \cdot (C/\log N)^p$, with probability at least $1-\delta$, $\texttt{Cov}_N(\widehat{\pi}) \lesssim \frac{1}{\log N}\left(\frac{C^p}{n}\right)^{\frac{1}{p+1}} + \frac{\log(1/\delta)}{n}$.*

This result shows that for sufficiently rich classes (e.g., when $p > 0$), the fine-grained term dominates the coarse-grained term for n sufficently large. On the other hand, for simple classes (e.g., when $p = 0$), the coarse-grained term can dominate the fine-grained term.

Autoregressive linear models: Low dimension. We now specialize to our running example, the autoregressive linear model in Eq. (2). This class satisfies $\log \mathcal{N}_\infty(\Pi, \alpha) \asymp d \log(BH/\alpha)$ (corresponding to the parametric regime in Corollary 4.1), and so, coverage generalizes in a (nearly) horizon-independent fashion, in stark contrast to the cross-entropy lower bound in Proposition 3.2. The only drawback (which is fundamental) is that since the class has low capacity, the coarse-grained term dominates for most parameter regimes, and the improvement as N scales is quite modest.

Autoregressive linear models: High dimension. As a more interesting example, we next look at the behavior of next-token prediction for autoregressive linear models in an "overparameterized" regime where the dimension d is arbitrarily large (Zhang, 2002; Neyshabur et al., 2015; Bartlett et al., 2017); here we expect polynomial dependence on the norm parameter B, as it is the only parameter

that controls the richness of the class Π. In this regime, it turns out that in the worst-case, the capacity $\log \mathcal{N}_\infty(\Pi, \alpha)$ scales polynomially in H. To address, this we prove a refined version of Theorem 4.1 that adapts to the variance in the data distribution π_{D}, avoiding explicit dependence on sequence length.

Define the *inherent variance* for the data distribution as

$$\sigma_\star^2 := \mathbb{E}_{\pi_{\mathsf{D}}}\left[\sum_{h=1}^{H} \left\|\phi(x, y_{1:h}) - \bar{\phi}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\right\|^2\right], \tag{5}$$

where $\bar{\phi}_{\pi_{\mathsf{D}}}(x, y_{1:h-1}) := \mathbb{E}_{y_h \sim \pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1})}[\phi(x, y_{1:h})]$ is the average feature vector given the prefix $(x, y_{1:h-1})$. We can interpret the inherent variance $\sigma_\star^2$ as a notion of effective sequence length; it captures the number tokens that are "pivotal" in the sense that they have high variation conditioned on the prefix; the name reflects the observed phenomenon that, in language modeling, most tokens are near-deterministic given their prefix, with only a few having high entropy (Abdin et al., 2024). Thus, while $\sigma_\star^2$ can be as large as $B^2 H$ in the worst case, we expect it to be smaller in general.

Theorem 4.2 (Overparameterized autoregressive linear models)**.** *Consider the autoregressive linear model (2), and suppose Assumptions 2.1 and 2.2 hold. For any $N \geq 2$, next-token prediction achieves*

$$\mathbb{E}[\mathsf{Cov}_N(\widehat{\pi})] \lesssim \sqrt{\frac{\sigma_\star^2}{n \cdot \log N}} + \frac{B^2}{n}. \tag{6}$$

Similar to Theorem 4.1, the first term in Eq. (6) can be viewed as "fine-grained", but decreases with the tail parameter N, while the second "coarse-grained" term does not decrease with N but will typically be smaller to begin with. We prove (details in Proposition K.1) that this result is tight in the sense that if $\sigma_\star^2 \asymp H$, $n \geq H$ is indeed necessary to achieve good coverage in the high-dimensional regime.

We view the introduction of the inherent variance $\sigma_\star^2$ as an instance-dependent notion of complexity for autoregressive models to be a non-trivial conceptual contribution, which may find broader use.

5 Stochastic Gradient Descent Through the Lens of Coverage

The coverage-based generalization guarantees for next-token prediction in the prequel apply to general model classes Π, but consider the empirical maximizer $\widehat{\pi} = \arg\max_{\pi \in \Pi} \widehat{L}_n(\pi)$ of the next-token prediction (maximum likelihood) objective, in the vein of classical techniques in learning theory. For our second set of main results, we focus on autoregressive linear models (2) but take a more realistic approach and analyze stochastic gradient descent (SGD) in the single-pass regime. This setup is motivated by contemporary ("compute-optimal") language model training, which typically uses one or fewer passes over the training corpus (Kaplan et al., 2020; Hoffmann et al., 2022).

5.1 Stochastic Gradient Descent has Suboptimal Coverage

For the next-token prediction objective, single-pass stochastic gradient descent (SGD) takes the form[4]

$$\theta^{t+1} \leftarrow \mathrm{Proj}_\Theta(\theta^t + \eta \nabla \log \pi_{\theta^t}(y^t \mid x^t)), \tag{7}$$

for $x^t \sim \mu$ and $y^t \sim \pi_{\mathsf{D}}(\cdot \mid x^t)$, where $\eta > 0$ is the learning rate. As the next-token prediction loss $L(\theta) := \mathbb{E}_{\pi_{\mathsf{D}}}[-\log \pi_\theta(y \mid x)]$ is convex under this parameterization, we can show that SGD converges to π_{D} in KL divergence. This implies a coverage bound, albeit a suboptimal one.

Proposition 5.1 (SGD for autoregressive linear models)**.** ***Upper bound:*** *Suppose Assumptions 2.1 and 2.2 hold. As long as $\eta \leq \frac{1}{16HB^2}$, it holds that $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \pi_{\theta^t})\right] \leq \frac{1}{\eta T} + 4\eta\sigma_\star^2$. Choosing η to minimize this bound gives*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathsf{Cov}_N(\pi_{\theta^t})\right] \lesssim \frac{1}{\log N} \cdot \left(\sqrt{\frac{\sigma_\star^2}{T}} + \frac{B^2 H}{T}\right). \tag{8}$$

Lower bound: *Suppose that $B \geq c \cdot \log^2(TH)$. Then there exists an autoregressive linear class Π such that for any constant step size $\eta > 0$, there exists an instance $\pi_{\mathsf{D}} \in \Pi$ with $\sigma_\star \leq 1$ such that with probability at least 0.5, the SGD iterates satisfy $\mathsf{Cov}_N(\pi_{\mathsf{D}} \,\|\, \pi_{\theta^t}) \geq c \cdot \min\left\{\frac{H}{T \log N}, 1\right\}$ for any $t \in [T]$.*

The coverage bound in Eq. (8) (which follows by passing from KL to coverage through Proposition 3.1) is similar to Theorem 4.2, except that the second term $\frac{B^2 H}{T}$ has an unfortunate dependence on the sequence length H. The lower bound shows that this dependence is tight, and SGD can indeed

[4] $\mathrm{Proj}_\Theta(\cdot)$ denotes Euclidean projection onto Θ, so this is SGD on the loss $L(\theta) := \mathbb{E}[-\log \pi_\theta(y \mid x)]$.

experience poor coverage. The failure of SGD in Proposition 5.1 is related to *heterogeneity* across prompts: there are some prompts for which the effective scale of the gradient in Eq. (7) grows with H, leading to divergence unless we use a small learning rate $\eta \lesssim \frac{1}{HB}$. Yet for other prompts, the effective gradient range is small, leading to slow convergence (on the order of $\Omega(H)$ steps) unless $\eta \gg \frac{1}{HB}$.

Remark 5.1 (Sequence-level SGD)**.** *The update in Eq. (7) can be interpreted as a "sequence-level" form of SGD, since we perform a single gradient step for each full sequence y^t (note that $\nabla \log \pi_{\theta^t}(y^t \mid x^t) = \sum_{h=1}^{H} \nabla \log \pi_{\theta^t}(y_h^t \mid x^t, y_{1:h-1}^t)$). We view this as a model for what is done in practice, whereby one performs SGD on sequences of tokens spanning some fixed context window. While this context window may be shorter than the full training example (e.g., a long article), understanding the implications of a limited context window is beyond the scope of this work.*

5.2 Gradient Normalization Improves Coverage

To address the suboptimality of SGD, we consider *gradient normalization* as a simple intervention. For a mini-batch $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^{K}$ of K samples from π_{D}, define the batch stochastic gradient as $\widehat{g}(\theta; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \nabla \log \pi_\theta(y \mid x)$. We consider the following normalized SGD update:

$$\theta^{t+1} \leftarrow \mathrm{Proj}_\Theta\big(\theta^t + \eta \cdot \tfrac{\widehat{g}(\theta^t; \mathcal{D}^t)}{\lambda + \|\widehat{g}(\theta^t; \mathcal{D}^t)\|}\big); \tag{9}$$

here $\mathcal{D}^t$ is a mini-batch with K fresh samples drawn i.i.d. from π_{D}, and $\lambda > 0$ is a regularization parameter for numerical stability. We show that this update achieves a horizon-independent coverage bound.

Theorem 5.1. *Suppose Assumption 2.1 and Assumption 2.2 hold. Let $T, K \geq 1, N \geq 3$ be given. For an appropriate choice of $\eta, \lambda > 0$, the normalized SGD update (9) achieves the following bound:*

$$\mathbb{E}\Big[\tfrac{1}{T}\textstyle\sum_{t=1}^{T} \mathsf{Cov}_N(\pi_{\theta^t})\Big] \lesssim \sqrt{\tfrac{\sigma_\star^2}{T \cdot \log N}} + \tfrac{B^2}{T} + \tfrac{B}{K \cdot \log N}. \tag{10}$$

To achieve $\mathbb{E}[\mathsf{Cov}_N(\widehat{\pi})] \leq \varepsilon$ for a target level $\varepsilon > 0$, it suffices to choose $T = O\big(\frac{\sigma_\star^2}{\varepsilon^2 \log N} + \frac{B^2}{\varepsilon}\big)$, $K = O\big(\frac{B}{\varepsilon \log N} + 1\big)$, giving total sample complexity $n = TK = O\big(\frac{\sigma_\star^2 B}{\varepsilon^3 \log^2 N} + \frac{B^3 + \sigma_\star^2}{\varepsilon^2 \log N} + \frac{B^2}{\varepsilon}\big)$.

Theorem 5.1 shows that gradient normalization achieves horizon-independent coverage with a qualitatively similar rate to the guarantee for next-token prediction in Theorem 4.2: To achieve coverage ε, both rates scale as $\mathrm{poly}\big(\frac{\sigma_\star^2}{\log N}, B, \varepsilon^{-1}\big)$, though the dependence on ε for Theorem 5.1 is worse. We view this as another instance of the coverage principle, as the rate achieved by gradient normalization goes beyond what can be achieved by passing through KL divergence. We emphasize that minibatching alone is not enough to achieve this result; rather, minibatching is necessary to avoid excessive bias once we introduce gradient normalization.

As a remark, the normalized SG update in (9) is closely related to SignSGD (Balles & Hennig, 2018) and *Adam* (Kingma & Ba, 2015) as shown by Bernstein & Newhouse (2024). We believe that similar coverage guarantees could potentially be shown for these methods using our techniques.

Distillation. As an additional result, we show (Theorem F.2 in Appendix F.3) that for a *distillation* setting, where π_{D} corresponds to a teacher model and we have access to its per-token logits, we can derive an improved gradient normalization scheme that fully closes the gap with Theorem 4.2.

6 Interventions for Better Coverage

In this section, we develop new interventions that improve coverage (and downstream performance) beyond the conventional algorithms analyzed in Sections 4 and 5. We view these results as promising proofs of concept, opening the door for further research into interventions driven by coverage.

6.1 Improving Coverage at Test Time

In this section, we show that a new test-time decoding strategy inspired by *test-time training* (Sun et al., 2020; Akyürek et al., 2025) leads to improved coverage when combined with token-level SGD.

We begin by departing from Eq. (7) and learning models with a *token-level* SGD update, defined as

$$\theta^{t,h+1} = \mathrm{Proj}_\Theta\big(\theta^{t,h} + \eta \nabla \log \pi_{\theta^{t,h}}(y_h^t \mid x^t, y_{1:h-1}^t)\big), \quad \text{for } h = 0, \cdots, H-1, \tag{11}$$

and $\theta^{t+1} \equiv \theta^{t+1,0} := \theta^{t,H}$ for $t \in [T]$, and where $(x^t, y^t_{1:H}) \sim \pi_{\mathsf{D}}$. Below we show that, when combined with a test-time training-like update that performs token-level gradient updates *during test time*, the updates in Eq. (11) can circumvent the H-dependence in the lower bound of Proposition 5.1.

Concretely, we consider a distribution $\pi^{\mathsf{TTT}}_\theta : \mathcal{X} \to \Delta(\mathcal{Y}^H)$ formally introduced in Appendix L.1, which can be interpreted as an augmented version of the autoregressive linear model π_θ that uses test-time training to sample. Given a prompt x, we first sample $y_1 \sim \pi_\theta(\cdot \mid x)$, then perform a gradient step $\theta' \leftarrow \mathrm{Proj}_\Theta(\theta + \eta \nabla \log \pi_\theta(y_1 \mid x))$ to increase the probability of the token we just sampled. We then sample $y_2 \sim \pi_{\theta'}(\cdot \mid x, y_1)$, update $\theta'' \leftarrow \mathrm{Proj}_\Theta(\theta' + \eta \nabla \log \pi_{\theta'}(y_2 \mid x, y_1))$, and so on. Once the full sequence $y_{1:H}$ is sampled, we reset back to θ, to process the next example at test-time. We show that when augmented with this test-time sampling scheme, token-level SGD achieves a horizon-independent coverage bound that matches and even slightly improves upon the bound for next-token prediction in Theorem 4.2.

Theorem 6.1 (Token-level SGD with test-time training)**.** *Suppose Assumption 2.1 and Assumption 2.2 hold. For a suitably chosen parameter $\eta > 0$, token-level SGD (11) achieves $\mathbb{E}\big[\frac{1}{T}\sum_{t=1}^T D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \pi^{\mathsf{TTT}}_{\theta^t})\big] \lesssim \sqrt{\frac{\sigma_\star^2}{T} + \frac{B^2}{T}}$, and thus $\mathbb{E}\big[\frac{1}{T}\sum_{t=1}^T \mathsf{Cov}_N(\pi^{\mathsf{TTT}}_{\theta^t})\big] \lesssim \frac{1}{\log N}\big(\sqrt{\frac{\sigma_\star^2}{T} + \frac{B^2}{T}}\big)$.*

This improves Theorem 4.2 by a factor of $1/\sqrt{\log N}$ on the leading term and a factor of $1/\log N$ on the second term. Furthermore, the algorithm bypasses the lower bound on KL divergence for *proper* methods (Proposition 3.2), demonstrating a provable benefit of being *improper*.

6.2 Selecting for Coverage

We now consider the problem of selecting a model (e.g., checkpoint) from a small number of candidates to achieve the best coverage. We introduce a tournament-like procedure that improves upon maximum likelihood in that it removes the requirement that $\pi_{\mathsf{D}} \in \Pi$; it is guaranteed to find a model in the class with good coverage if one exists, even if π_{D} itself is not in the class. As an algorithmic intervention, we envision using this procedure to select a single training checkpoint or hyperparameter configuration to use for RL fine-tuning or test-time scaling. Indeed, as demonstrated in Figure 1, using cross-entropy as a selection criterion—as is standard—may result in poor coverage, and these procedures can be used to select better checkpoints. Our results here concern the general setting in Section 2, and are not restricted to autoregressive linear models.

A simple tournament estimator for coverage. Given a dataset $\mathcal{D} = \{(x^i, y^i)\}_{i \in [n]}$, define

$$\widehat{\mathsf{Cov}}_N(\pi' \,\|\, \pi) := \frac{1}{n}\Big|\Big\{i \in [n] : \frac{\pi'(y^i|x^i)}{\pi(y^i|x^i)} \geq N\Big\}\Big|, \tag{12}$$

which can be interpreted as an empirical version of the coverage profile $\mathsf{Cov}_N(\pi' \,\|\, \pi)$ in Eq. (1) when $\pi' = \pi_{\mathsf{D}}$. For $N \geq 1$, we consider the estimator

$$\widehat{\pi} := \arg\min_{\pi \in \Pi} \max_{\pi' \in \Pi} \widehat{\mathsf{Cov}}_N(\pi' \,\|\, \pi). \tag{13}$$

Intuitively, this estimator chooses the model π that minimizes the maximum coverage against any other model π' in the class Π. When Π is small, we can implement this tournament by simply evaluating the empirical coverage in Eq. (12) for each pair. The main guarantee for this estimator is as follows.

Theorem 6.2. *Let $N \geq 1$ be given. Then, for any $a \in [0, 1]$, with probability at least $1 - \delta$, the tournament estimator (13) achieves*

$$\mathsf{Cov}_{N^{1+a}}(\widehat{\pi}) \lesssim \min_{\pi \in \Pi} \mathsf{Cov}_{N^a}(\pi) + \frac{1}{N^{1-a}} + \frac{\log(|\Pi|/\delta)}{n}. \tag{14}$$

This shows that the tournament achieves a coverage profile nearly as good as the best-in class, except for a small polynomial blow up, in that we bound the coverage at level N^{1+a} in terms of the coverage for the best-in class at level N^a.

Infinite class and improving the tournament. Eq. (13) can also be applied to general, infinite classes Π. In this case, it turns out that it improves upon the coverage achieved by the maximum likelihood estimator in Theorem 4.1 (see Theorem 6.2′). Furthermore, in Appendix F.4, we describe an improved tournament estimator that is able to remove the $1/N^{1-a}$ term from Theorem 6.2, thereby achieving nontrivial guarantees even when the coverage parameter N is a constant.

Discussion and Future Work

See Appendix A for discussion and open problems, and Appendix F for additional results.

Reproducibility Statement

We provide full proofs for all theoretical results in the appendix. Appendix G includes extensive experiment setup and implementation details for all empirical results. The source code is included in the supplementary material, along with the plotting scripts and data to reproduce Figure 1 and Figure 2.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Ekin Akyürek, Mehul Damani, Adam Zweiger, Linlu Qiu, Han Guo, Jyothish Pari, Yoon Kim, and Jacob Andreas. The surprising effectiveness of test-time training for few-shot learning. In *Forty-second International Conference on Machine Learning*, 2025.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FxNNiUgtfa.

Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. *arXiv preprint arXiv:2403.06963*, 2024.

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pp. 404–413. PMLR, 2018.

Hritik Bansal, Arian Hosseini, Rishabh Agarwal, Vinh Q. Tran, and Mehran Kazemi. Smaller, weaker, yet better: Training LLM reasoners via compute-optimal sampling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3OyaXFQuDl.

Peter L. Bartlett and Andrea Montanari. Deep learning: A statistical viewpoint. *Acta Numerica*, 30: 87–201, 2021.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences (PNAS)*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences (PNAS)*, 116(32):15849–15854, 2019.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: Optimization for Machine Learning*, 2024.

Blair Bilodeau, Dylan J Foster, and Daniel M Roy. Minimax rates for conditional density estimation via empirical entropy. *The Annals of Statistics*, 2023.

Adam Block and Yury Polyanskiy. The sample complexity of approximate rejection sampling with applications to smoothed online learning. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 228–273. PMLR, 2023.

Bradley Brown, Jordan Juravsky, Ryan Saul Ehrlich, Ronald Clark, Quoc V Le, Christopher Re, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2025. URL https://openreview.net/forum?id=0xUEBQV54B.

Feng Chen, Allan Raventos, Nan Cheng, Surya Ganguli, and Shaul Druckmann. Rethinking fine-tuning when scaling test-time compute: Limiting confidence improves mathematical reasoning. *arXiv preprint arXiv:2502.07154*, 2025.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pp. 1042–1051. PMLR, 2019.

Yangyi Chen, Binxuan Huang, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. Scaling laws for predicting downstream performance in llms. *Transactions on Machine Learning Research*, 2024.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dYur3yabMj.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 2010.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Marc Finzi, Sanyam Kapoor, Diego Granziol, Anming Gu, Christopher De Sa, J Zico Kolter, and Andrew Gordon Wilson. Compute-optimal llms provably generalize better with scale. *arXiv preprint arXiv:2504.15208*, 2025.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv:2112.13487*, 2021.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory*, pp. 3489–3489. PMLR, 2022.

Dylan J Foster, Adam Block, and Dipendra Misra. Is behavior cloning all you need? understanding horizon in imitation learning. *arXiv preprint arXiv:2407.15007*, 2024.

Dylan J Foster, Zakaria Mhammedi, and Dhruv Rohatgi. Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration. *Conference on Learning Theory (COLT)*, 2025.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, et al. Language models scale reliably with over-training and on downstream tasks. In *The Thirteenth International Conference on Learning Representations*, 2024.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Luca Soldaini, Jenia Jitsev, Alex Dimakis, Gabriel Ilharco, Pang Wei Koh, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. Language models scale reliably with over-training and on downstream tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iZeQBqJamf.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

Zhaolin Gao, Jonathan D Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, J Andrew Bagnell, Jason D Lee, and Wen Sun. REBEL: Reinforcement learning via regressing relative rewards. *arXiv:2404.16767*, 2024.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. Scaling laws for neural machine translation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hR_SMu8cxCV.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 30016–30030, 2022.

Audrey Huang, Adam Block, Dylan J Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. *International Conference on Learning Representations (ICLR)*, 2025a.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. *International Conference on Machine Learning (ICML)*, 2025b.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025c.

Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. Compression represents intelligence linearly. In *First Conference on Language Modeling*, 2024.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Xiang Ji, Sanjeev Kulkarni, Mengdi Wang, and Tengyang Xie. Self-play with adversarial critic: Provable and scalable offline alignment for language models. *arXiv:2406.04274*, 2024.

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 2024.

Hangzhan Jin, Sitao Luan, Sicheng Lyu, Guillaume Rabusseau, Reihaneh Rabbany, Doina Precup, and Mohammad Hamdaqa. Rl fine-tuning heals ood forgetting in sft, 2025. URL `https://arxiv.org/abs/2509.12235`.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International conference on machine learning*, pp. 5084–5096. PMLR, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same Pre-training Loss, Better Downstream: Implicit Bias Matters for Language Models, 2022.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in RLHF: Your SFT loss is implicitly an adversarial regularizer. *arXiv:2405.16436*, 2024.

Sanae Lotfi, Marc Finzi, Yilun Kuang, Tim GJ Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. *arXiv preprint arXiv:2312.17173*, 2023.

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

Nicholas Lourie, Michael Y Hu, and Kyunghyun Cho. Scaling laws are unreliable for downstream tasks: A reality check. *arXiv preprint arXiv:2507.00885*, 2025.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuanjing Huang, and Xipeng Qiu. Scaling laws for fact memorization of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11263–11282, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.658. URL https://aclanthology.org/2024.findings-emnlp.658/.

Shahar Mendelson. Learning without Concentration. In *Conference on Learning Theory*, 2014.

Shahar Mendelson. Extending the scope of the small-ball method. *arXiv preprint arXiv:1709.00843*, 2017.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Vaishnavh Nagarajan, Chen Henry Wu, Charles Ding, and Aditi Raghunathan. Roll the dice & look before you leap: Going beyond the creative limits of next-token prediction. In *Forty-second International Conference on Machine Learning*, 2025.

Behnam Neyshabur, Ryota Tomioka, and Nati Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory (COLT)*, 2015.

OpenAI. Introducing openai o1. *Blog*, 2024. URL https://openai.com/o1/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alexander Rakhlin and Karthik Sridharan. Statistical learning and sequential prediction, 2012. Available at http://www.mit.edu/~rakhlin/courses/stat928/stat928_notes.pdf.

Dhruv Rohatgi, Adam Block, Audrey Huang, Akshay Krishnamurthy, and Dylan J. Foster. Computational-statistical tradeoffs at the next-token prediction barrier: Autoregressive and imitation learning under misspecification. *arXiv preprint arXiv:2502.12465*, 2025.

Clayton Sanford, Bahare Fatemi, Ethan Hall, Anton Tsitsulin, Mehran Kazemi, Jonathan Halcrow, Bryan Perozzi, and Vahab Mirrokni. Understanding transformer reasoning capabilities via graph algorithms. *Advances in Neural Information Processing Systems*, 37:78320–78370, 2024.

Abulhair Saparov, Srushti Ajay Pawar, Shreyas Pimpalgaonkar, Nitish Joshi, Richard Yuanzhe Pang, Vishakh Padmakumar, Mehran Kazemi, Najoung Kim, and He He. Transformers struggle to learn to search. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=0bmXrtTDUu.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Yuda Song, Gokul Swamy, Aarti Singh, J Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. *Advances in Neural Information Processing Systems*, 37:12243–12270, 2024.

Vladimir Spokoiny. Parametric estimation. finite sample theory. *The Annals of Statistics*, pp. 2877–2909, 2012.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=YW6edSufht.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Jianheng Tang, Qifan Zhang, Yuhan Li, Nuo Chen, and Jia Li. Grapharena: Evaluating and exploring large language models on graph computation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alexander K Taylor, Anthony Cuturrufo, Vishal Yathish, Mingyu Derek Ma, and Wei Wang. Are large-language models graph algorithmic reasoners? *arXiv preprint arXiv:2410.22597*, 2024.

S. A. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2000.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? *Advances in Neural Information Processing Systems*, 36:30840–30861, 2023.

Xinyi Wang, Shawn Tan, Mingyu Jin, William Yang Wang, Rameswar Panda, and Yikang Shen. Do larger language models imply better generalization? a pretraining scaling law for implicit reasoning, 2025. URL https://arxiv.org/abs/2504.03635.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, 1995.

Fang Wu, Weihao Xuan, Ximing Lu, Zaid Harchaoui, and Yejin Choi. The Invisible Leash: Why RLVR May Not Escape Its Origin, 2025.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. Training trajectories of language models across scales. *arXiv preprint arXiv:2212.09803*, 2022.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Yuhong Yang and Andrew R Barron. An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44(1):95–116, 1998.

Gilad Yehudai, Noah Amsel, and Joan Bruna. Compositional reasoning with transformers, rnns, and chain of thought. *arXiv preprint arXiv:2503.01544*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Hansi Zeng, Kai Hui, Honglei Zhuang, Zhen Qin, Zhenrui Yue, Hamed Zamani, and Dana Alon. Can Pre-training Indicators Reliably Predict Fine-tuning Outcomes of LLMs?, 2025.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Tong Zhang. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 2006.

CONTENTS OF APPENDIX

Part I

Additional Discussion and Results

A Discussion and Future Work

Our work, through the lens of coverage, takes a first step toward clarifying the mechanisms through which pre-training with next-token prediction leads to models for which post-training is effective.

A.1 Simplifications in the Problem Formulation

In the course of the paper we have made various simplifying assumptions, some of which can be relaxed in a straightforward fashion, while others are more fundamental:

- In language model pre-training, the pre-training corpus consists of sequences y with varying lengths H, and does not typically split examples into prompts and responses. Our formulation in Section 2 is a simplification (one that is closer in spirit to supervised fine-tuning), but we expect that the insights derived here can extend to the general setting.
- Much of our analysis focuses on the realizable/well-specified setting where $\pi_{\mathsf{D}} \in \Pi$. We give evidence in Appendix F that the coverage profile is more tolerant to misspecification than KL-divergence, but we leave a deeper investigation for future work.
- Our treatment assumes the distribution over prompts μ is the same for pre-training and post-training. This is straightforward to relax at the cost of introducing an additional coverage or distribution shift coefficient to handle the mismatch between the two distributions.
- We show that a good coverage profile is necessary for BoN to succeed on downstream tasks. While there is ample evidence current RL techniques can fail in the absence of coverage (Yue et al., 2025; Gandhi et al., 2025; Wu et al., 2025), it is not clear what the minimal conditions required for RL are, and they may be weaker than coverage.

A.2 Future Work

Our results open several new directions for future research.

- *Interventions for coverage.* There is much to be done in understanding and improving existing algorithms such as optimizers through the lens of coverage. Our results in Section 6 show initial promise for using coverage to guide design of optimizers and model selection schemes, but the algorithm design space remains opaque, and there may be significant room for futher improvement. More ambitiously, one could imagine re-structuring the entire language modeling pipeline itself around coverage.
- *Semantic coverage.* The notion of coverage we focus on, the *coverage profile*, is mathematically convenient but may be conservative in regard to downstream performance, since it only depends on the model through its predicted probabilities. An important direction for future work is to understand pre-training and post-training through fine-grained "semantic" notions of coverage that more explicitly account for the representations learned by next-token prediction.

B Additional Related Work

Related empirical observations. On the empirical side, our results are connected to a line of work that studies to scaling laws for zero-shot downstream performance based on pre-training metrics such as cross-entropy (Gadre et al., 2024; Huang et al., 2024; Chen et al., 2024; Sardana et al., 2024). Several empirical works have also investigated how specific capabilities scale with additional pre-training, including machine translation (Ghorbani et al., 2022), knowledge capacity and memorization (Allen-Zhu & Li, 2025; Lu et al., 2024), and multi-hop reasoning (Wang et al., 2025). Our findings are consistent with Liu et al. (2022); Zeng et al. (2025); Lourie et al. (2025); Springer et al. (2025), who observe that cross-entropy is not always sufficient for predicting downstream performance, and in some cases can be anti-correlated.

Perhaps most closely related, Chen et al. (2025) show empirically the decreasing cross-entropy in pre-training does not necessarily lead to better pass@N performance, and that pass@N can even degrade as pre-training proceeds—a finding similar to Figure 1.[5] Our theoretical results can be viewed as placing their findings on stronger theoretical footing; conversely, their empirical results provide strong motivation for our theoretical treatment. Chen et al. (2025) also study a modification to the maximum likelihood objective aimed at improving coverage (in the spirit of Section 6), but, when instantiated with chain-of-thought, their approach requires a small space of possible final answers.

We mention in passing a few additional works. Chu et al. (2025) explored the different synergistic roles that supervised fine-tuning (SFT) and RL play in language model development, and subsequent work observed that the best checkpoint to start RL from can sometimes be in the middle of SFT (Jin et al., 2025). Bansal et al. (2025) empirically identified the coverage of teacher-generated synthetic data as an important indicator of how effective distillation would be for reasoning tasks. Several papers have also investigated empirical tradeoffs between model size and reasoning performance under best-of-N sampling (Snell et al., 2025; Brown et al., 2025).

Coverage in post-training. Coverage metrics similar to coverage profile play a central role in theoretical literature on post-training and test-time algorithms (Huang et al., 2025a;b;c; Foster et al., 2025; Liu et al., 2024; Song et al., 2024; Gao et al., 2024; Liu et al., 2024; Ji et al., 2024), which analyze algorithms under the assumption that the base model has good coverage; our work can be viewed as providing theoretical motivation for this assumption.

Various notions of coverage similar to coverage profile have also appeared in the more classical literature on offline reinforcement learning (Farahmand et al., 2010; Chen & Jiang, 2019; Xie & Jiang, 2020; Jin et al., 2021; Foster et al., 2022; Jiang & Xie, 2024); here coverage is typically used to quantify the quality of an offline dataset rather than a model/policy itself.

Generalization in deep learning. Understanding the generalization behavior of deep learning models has been a central focus of the theory community for the last decade (Neyshabur et al., 2015; Zhang et al., 2017; Bartlett et al., 2017; Jacot et al., 2018; Belkin et al., 2019; Nagarajan & Kolter, 2019; Bartlett et al., 2020; Bartlett & Montanari, 2021). Our approach is somewhat complementary, in the sense that it focuses on the specific objective of next-token prediction with the logarithmic loss, and aims to understand when minimizing this loss leads to generalization for an *alternative* objective, coverage profile. We expect that our techniques can be combined with these more general results to provide more refined understanding of generalization for coverage profile with deep models.

From this line of work, perhaps most closely related are Lotfi et al. (2023; 2024); Finzi et al. (2025), which aim to provide non-vacuous generalization bounds for the cross-entropy loss itself for autoregressive models.

Analysis of maximum likelihood. On the theoretical side, our results are most closely related to a classical line of work in statistics (Wong & Shen, 1995; van de Geer, 2000; Zhang, 2006), which shows that maximum likelihood can converge to the true model in Hellinger distance (or other Renyi divergences) under minimal assumptions, even when KL divergence is poorly behaved (large or infinite). Our results in Section 4 are similar in spirit, but provide a more fine-grained perspective, showing that coverage profile can converge even faster than these results might suggest, particularly as one ventures further into the tail. Our analysis has some conceptual similarity to the small ball method of Mendelson (2014; 2017), which we elaborate on in Appendix J.1.

Our techniques are also related to recent work of Foster et al. (2024); Rohatgi et al. (2025), which specialize the general techniques above to autoregressive models (e.g., under Hellinger distance).

C Properties of the Coverage Profile

Before proceeding, we briefly discuss some conceptual properties of the coverage profile that will be helpful to keep in mind.

Remark C.1 (Coverage profile as a refinement of cross-entropy)**.** *While we position the coverage profile as a new quantity of interest, it can also be viewed as a fine-grained, inference budget-sensitive*

[5] While Chen et al. (2025) use the term "coverage", it is used as a synonym for pass@N, and is not specifically related to the notion of the coverage profile we consider here.

refinement *of cross-entropy. Concretely, if we write*

$$\mathsf{Cov}_N(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) = \mathbb{P}_{\pi_{\mathsf{D}}}\left[\log \frac{\pi_{\mathsf{D}}(y \mid x)}{\widehat{\pi}(y \mid x)} \geq \log N\right], \tag{15}$$

it becomes clear that the coverage profile is simply the cumulative distribution function (CDF) of the log density ratio $X := \log \frac{\pi_{\mathsf{D}}(y|x)}{\widehat{\pi}(y|x)}$*, while KL-divergence corresponds to the mean:* $\mathbb{E}_{\pi_{\mathsf{D}}}[X]$*. It is well known that the CDF of a random variable is a more informative statistic than its mean (Durrett, 2019); the former can be much more sensitive to the model's behavior at the tail than the latter. Indeed, the coverage profile can behave very differently across scales, as shown by Figure 1.*[6]

Remark C.2 (KL divergence and coverage profile are not estimable)**.** *We emphasize that KL-divergence and the coverage profile are not estimable quantities in general, due to the fact both depend on the unknown density* $\pi_{\mathsf{D}}(y \mid x)$ *for the data distribution. This motivates the use of cross-entropy in practice, as the former is an estimable upper bound on* $D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi})$*. Analogously, we show in Section 6.2 that various estimable proxies for the coverage profile can be used to select models with good coverage.*

An exception is the expert distillation *setting (see Section 6.1), where* π_{D} *is a teacher network for which the log-probabilities* $\log \pi_{\mathsf{D}}(y \mid x)$ *are available.*

Remark C.3 (Sequence-level versus answer-level coverage)**.** *Our discussion so far has focused on coverage at the* sequence level*. For reasoning tasks, it is natural to explicitly factorize the response* $y = (y_{\mathsf{cot}}, y_{\mathsf{ans}})$ *into a chain-of-thought (reasoning trajectory) component* y_{cot} *and an answer component* y_{ans}*. For this setting, a weaker notion coverage is the following* answer-level coverage profile*:*

$$\mathsf{Cov}^{\mathsf{ans}}_N(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) := \mathbb{P}_{\pi_{\mathsf{D}}}\left[\frac{\pi_{\mathsf{D}}(y_{\mathsf{ans}} \mid x)}{\widehat{\pi}(y_{\mathsf{ans}} \mid x)} \geq N\right].$$

Informally, the answer-level coverage profile is sufficient for downstream BoN success for tasks where it is only important to produce the right answer, not a correct reasoning trace. We have $\mathsf{Cov}^{\mathsf{ans}}_N(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) \leq \mathsf{Cov}_N(\pi_{\mathsf{D}} \,\|\, \widehat{\pi})$*, but the former can be strictly smaller in general.*

We hope that by providing a comprehensive understanding of sequence-level coverage, our work can set the stage for future research on answer-level coverage and other finer-grained notions of coverage; we give some initial results along these lines in Appendix F.

[6] Interestingly, we show (Proposition D.1) that if the coverage profile satisfies a certain growth condition uniformly for all scales M, then it implies a bound on KL-divergence—a weak converse to Proposition 3.1.

D SUPPORTING RESULTS

D.1 PROPERTIES OF THE COVERAGE PROFILE

Proposition D.1 (KL-to-coverage conversion). *For all models π_{D} and π and $M \geq 2$, we have*

$$\mathsf{Cov}_N(\pi) \leq \frac{D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \pi)}{\log N - 1 + \frac{1}{N}}.$$

Proof of Proposition D.1. Lemma 27 of Block & Polyanskiy (2023) states that for any $N > 1$ and any convex $f : [0, \infty] \to [0, \infty]$ with $f(1) = f'(1) = 0$,

$$\mathsf{Cov}_N(\pi) = \mathbb{P}_{\pi_{\mathsf{D}}}\left[\frac{\pi_{\mathsf{D}}(y \mid x)}{\pi(y \mid x)} > N\right] \leq \frac{N D_f(\pi_{\mathsf{D}} \,\|\, \pi)}{f(N)}, \tag{16}$$

where $D_f(\pi_{\mathsf{D}} \,\|\, \pi) := \mathbb{E}_\pi\left[f\left(\frac{\mathrm{d}\pi_{\mathsf{D}}}{\mathrm{d}\pi}\right)\right]$. Applying this with KL-divergence, which corresponds to $f(x) = x \log x - x + 1$ with $f'(x) = \log x$, we have that

$$\frac{N}{f(N)} = \frac{1}{\log N - 1 + 1/N}, \tag{17}$$

which gives the result.

□

Proposition D.2 (Tightness of KL-to-coverage conversion). *For any $N \geq 2$, there exist models π_{D} and $\widehat{\pi}$ such that*

$$\mathsf{Cov}_N(\widehat{\pi}) \geq \frac{D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi})}{\log N - \frac{1}{2} + \frac{1}{2N}}.$$

Proof of Proposition D.2. Consider $\pi_{\mathsf{D}} = \mathrm{Ber}(p)$ and $\widehat{\pi} = \mathrm{Ber}(p/N)$ with $p \leq \frac{1}{2}$. Then $\mathsf{Cov}_N(\widehat{\pi}) = p$ and

$$\begin{aligned} D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) &= p \log N + (1-p) \log \frac{1-p}{1-\frac{p}{N}} \leq p \log N + (1-p)\left(\frac{1-p}{1-\frac{p}{N}} - 1\right) \\ &= p\left(\log N - (1-p)\frac{1-\frac{1}{N}}{1-\frac{p}{N}}\right) \\ &\leq p \cdot \left(\log N - \frac{1}{2} + \frac{1}{2N}\right). \end{aligned}$$

This is the desired result. □

Proposition D.3 (Uniform coverage decay implies bounded KL). *Given $\pi, \pi_{\mathsf{D}} : \mathcal{X} \to \Delta(\mathcal{Y})$, define $W_{\mathsf{max}} := \sup_{x,y} \frac{\pi_{\mathsf{D}}(y|x)}{\pi(y|x)}$ and*

$$C := \sup_{N \geq 1} \{\mathsf{Cov}_N(\pi) \cdot \log N\},$$

where we note that $C \leq \log W_{\mathsf{max}}$. It holds that

$$D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \pi) \leq C \cdot (1 + \log(\log(W_{\mathsf{max}})/C)). \tag{18}$$

Proof of Proposition D.3. Let $\delta > 0$ a fixed parameter, and define $X := \pi_{\mathsf{D}}/\pi$. Then we have

$$D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \pi) = \mathbb{E}_{\pi_{\mathsf{D}}}[\log(X)] \leq \mathbb{E}_{\pi_{\mathsf{D}}}[\log(X)\mathbb{I}\{\log(X) > \delta\}] + \delta. \tag{19}$$

Since $X \leq W_{\max}$ almost surely, we can write

$$\mathbb{E}_{\pi_{\mathsf{D}}}[\log(X)\mathbb{I}\{\log(X) > \delta\}] = \int_{\delta}^{\log(W_{\max})} \mathbb{P}_{\pi_{\mathsf{D}}}[\log(X) > t]\mathrm{d}t \tag{20}$$

$$= \int_{\delta}^{\log(W_{\max})} \mathbb{P}_{\pi_{\mathsf{D}}}\big[X > e^t\big]\mathrm{d}t \tag{21}$$

$$\leq C \int_{\delta}^{\log(W_{\max})} \frac{1}{t}\mathrm{d}t \tag{22}$$

$$= C \log\left(\frac{\log(W_{\max})}{\delta}\right). \tag{23}$$

The result now follows by setting $\delta = C$. □

Proposition D.4 (Hellinger-to-coverage conversion)**.** *For all models π_{D} and π and $N > 1$, we have*

$$\mathsf{Cov}_N(\pi_{\mathsf{D}} \,\|\, \pi) \leq \frac{2N}{(\sqrt{N}-1)^2} \cdot D^2_{\mathsf{H}}(\pi_{\mathsf{D}}, \pi).$$

Proof of Proposition D.4. Without loss of generality, we assume $\mathcal{Y}$ is discrete in the following proof. By definition,

$$\begin{aligned}
D^2_{\mathsf{H}}(\pi_{\mathsf{D}}, \pi) &= \frac{1}{2}\,\mathbb{E}_{x\sim\pi_{\mathsf{D}}}\left[\sum_y \left(\sqrt{\pi_{\mathsf{D}}(y \mid x)} - \sqrt{\pi(y \mid x)}\right)^2\right] \\
&\geq \frac{1}{2}\,\mathbb{E}_{x\sim\pi_{\mathsf{D}}}\left[\sum_y \pi_{\mathsf{D}}(y \mid x)\left(1 - \frac{1}{\sqrt{N}}\right)^2 \mathbb{I}\left\{\pi(y \mid x) \leq \frac{1}{N}\pi_{\mathsf{D}}(y \mid x)\right\}\right] \\
&= \frac{1}{2}\left(1 - \frac{1}{\sqrt{N}}\right)^2 \mathbb{P}_{\pi_{\mathsf{D}}}\left[\frac{\pi_{\mathsf{D}}(y \mid x)}{\pi(y \mid x)} > N\right],
\end{aligned}$$

where the inequality follows from the fact that $\sqrt{\pi_{\mathsf{D}}(y \mid x)} - \sqrt{\pi(y \mid x)} \geq \left(1 - \frac{1}{\sqrt{N}}\right)\sqrt{\pi_{\mathsf{D}}(y \mid x)}$ is implied by $\pi(y \mid x) \leq \frac{1}{N}\pi_{\mathsf{D}}(y \mid x)$. Re-organizing completes the proof. □

Proposition D.5 (Chain rule for coverage profile)**.** *For any models π_{D}, π_{T}, and $\widehat{\pi}$, and any $M_1, M_2 \geq 2$, we have*

$$\mathsf{Cov}_{M_1}(\pi_{\mathsf{T}} \,\|\, \widehat{\pi}) \leq M_2 \cdot \mathsf{Cov}_{M_1/M_2}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) + \mathsf{Cov}_{M_2}(\pi_{\mathsf{T}} \,\|\, \pi_{\mathsf{D}}). \tag{24}$$

Proof of Proposition D.5. We can write

$$\begin{aligned}
\mathsf{Cov}_{M_1}(\pi_{\mathsf{T}} \,\|\, \widehat{\pi}) &= \mathbb{P}_{\pi_{\mathsf{T}}}\left[\frac{\pi_{\mathsf{T}}(y \mid x)}{\widehat{\pi}(y \mid x)} > M_1\right] \\
&= \mathbb{P}_{\pi_{\mathsf{T}}}\left[\frac{\pi_{\mathsf{T}}(y \mid x)}{\widehat{\pi}(y \mid x)} > M_1, \frac{\pi_{\mathsf{T}}(y \mid x)}{\pi_{\mathsf{D}}(y \mid x)} \leq M_2\right] + \mathbb{P}_{\pi_{\mathsf{T}}}\left[\frac{\pi_{\mathsf{T}}(y \mid x)}{\widehat{\pi}(y \mid x)} > M_1, \frac{\pi_{\mathsf{T}}(y \mid x)}{\pi_{\mathsf{D}}(y \mid x)} > M_2\right] \\
&\leq M_2 \mathbb{P}_{\pi_{\mathsf{D}}}\left[\frac{\pi_{\mathsf{D}}(y \mid x)}{\widehat{\pi}(y \mid x)} > M_1/M_2\right] + \mathbb{P}_{\pi_{\mathsf{T}}}\left[\frac{\pi_{\mathsf{T}}(y \mid x)}{\pi_{\mathsf{D}}(y \mid x)} > M_2\right] \\
&= M_2 \mathsf{Cov}_{M_1/M_2}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) + \mathsf{Cov}_{M_2}(\pi_{\mathsf{T}} \,\|\, \pi_{\mathsf{D}}).
\end{aligned}$$

□

D.2 Analysis of Best-of-N Sampling under a Good Coverage Profile

In this section we analyze the performance of the Best-of-N algorithm under a good coverage profile. Let a base model $\widehat{\pi}$ be given, and let a reward function $r_{\mathsf{T}}(x, y) \in [0, 1]$ be given. Let $\pi_{\mathsf{T}} : \mathcal{X} \to \Delta(\mathcal{Y})$ denote an arbitrary task-specific comparator policy.

We let $\widehat{\pi}^{\mathsf{BoN}}_N(x)$ denote the distribution of the Best-of-N algorithm with parameter N, which draws N responses $y^1, \ldots, y^N \overset{\text{i.i.d.}}{\sim} \widehat{\pi}(\cdot \mid x)$ and returns $y = \arg\max_{y_i} r_{\mathsf{T}}(x, y_i)$.

Proposition D.6 (Coverage implies success for BoN). *Let $M \geq 1$ be given. For any $\varepsilon > 0$, if $N \geq 2M \log(\varepsilon^{-1})$ and $\mathsf{Cov}_M(\pi_\mathsf{T} \,\|\, \widehat{\pi}) \leq \frac{1}{2}$, then we are guaranteed that*

$$\mathbb{E}_{x\sim\mu}\big[r_\mathsf{T}(x, \pi_\mathsf{T}(x)) - r_\mathsf{T}(x, \widehat{\pi}^{\mathsf{BoN}}_N(x))\big] \leq \mathsf{Cov}_M(\pi_\mathsf{T} \,\|\, \widehat{\pi}) + \varepsilon. \tag{25}$$

Proof of Proposition D.6. This is an immediate consequence of Lemma F.1 in Huang et al. (2025b), noting that we can bound $\mathcal{E}_M(\pi_\mathsf{T} \,\|\, \widehat{\pi}) \leq \mathsf{Cov}_M(\pi_\mathsf{T} \,\|\, \widehat{\pi})$. □

Proposition D.7 (Coverage is necessary for BoN). *For any model $\widehat{\pi}$ and reference π_T, and for any $N \geq 2$, there exists a reward function $r_\mathsf{T}(x, y) \in \{0, 1\}$ such that*

$$\mathbb{E}_{x\sim\mu}\big[r_\mathsf{T}(x, \pi_\mathsf{T}(x)) - r_\mathsf{T}(x, \widehat{\pi}^{\mathsf{BoN}}_N(x))\big] \geq \frac{1}{2}\mathsf{Cov}_{2N}(\pi_\mathsf{T} \,\|\, \widehat{\pi}). \tag{26}$$

Proof of Proposition D.7. For any $x \in \mathcal{X}$, we define $S_x := \big\{y \in \mathcal{Y} : \frac{\pi_\mathsf{T}(y|x)}{\widehat{\pi}(y|x)} \geq 2N\big\}$ and let $r_\mathsf{T}(x, y) = \mathbb{I}\{y \in S_x\}$.

By definition, for any fixed $x \in \mathcal{X}$, it holds that

$$\begin{aligned} r_\mathsf{T}(x, \widehat{\pi}^{\mathsf{BoN}}_N(x)) &= \mathbb{P}_{y\sim\widehat{\pi}^{\mathsf{BoN}}_N(x)}(y \in S_x) = \mathbb{P}_{y^1,\ldots,y^N \overset{\text{i.i.d.}}{\sim} \widehat{\pi}(\cdot|x)}(\exists i \in [N], y^i \in S_x) \\ &= 1 - \big(1 - \mathbb{P}_{y\sim\widehat{\pi}(\cdot|x)}(y \in S_x)\big)^N \leq N \cdot \mathbb{P}_{y\sim\widehat{\pi}(\cdot|x)}(y \in S_x) \\ &= N \cdot \sum_{y\in S_x} \widehat{\pi}(y \mid x) \leq N \cdot \sum_{y\in S_x} \frac{1}{2N}\pi_\mathsf{T}(y \mid x) = \frac{1}{2}\mathbb{P}_{y\sim\pi_\mathsf{T}(\cdot|x)}(S_x), \end{aligned}$$

where we use the fact that $\widehat{\pi}(y \mid x) \leq \frac{1}{2N}\pi_\mathsf{T}(y \mid x)$ for any $y \in S_x$. We also note that $\mathbb{P}_{x\sim\mu, y\sim\pi_\mathsf{T}(\cdot|x)}(y \in S_x) = \mathsf{Cov}_{2N}(\pi_\mathsf{T} \,\|\, \widehat{\pi})$. Therefore,

$$\mathbb{E}_{x\sim\mu}\big[r_\mathsf{T}(x, \pi_\mathsf{T}(x)) - r_\mathsf{T}(x, \widehat{\pi}^{\mathsf{BoN}}_N(x))\big] \geq \frac{1}{2}\mathsf{Cov}_{2N}(\pi_\mathsf{T} \,\|\, \widehat{\pi}).$$

□

D.3 Properties of Maximum Likelihood

Proposition D.8 (Convergence of maximum likelihood in Hellinger distance). *Assume that $\pi_\mathsf{D} \in \Pi$. With probability at least $1 - \delta$, the maximum likelihood estimator $\widehat{\pi} := \arg\max_{\pi\in\Pi} \widehat{L}_n(\pi)$ satisfies,*

$$D^2_\mathsf{H}(\pi_\mathsf{D}, \widehat{\pi}) \lesssim \inf_{\varepsilon>0}\left\{\frac{\log \mathcal{N}_\infty(\Pi, \varepsilon)}{n} + \varepsilon\right\}, \tag{27}$$

and consequently

$$\mathsf{Cov}_M(\widehat{\pi}) \lesssim \inf_{\varepsilon>0}\left\{\frac{\log \mathcal{N}_\infty(\Pi, \varepsilon)}{n} + \varepsilon\right\}. \tag{28}$$

for all $M \geq 2$.

Proof of Proposition D.8. The first bound follows from Proposition B.2 of Foster et al. (2024). The second bound follows from applying Proposition D.4. □

Proposition D.9 (Convergence of maximum likelihood in KL). *Assume that $\pi_\mathsf{D} \in \Pi$, and that all $\pi \in \Pi$ satisfy $\left\|\frac{\pi_\mathsf{D}}{\pi}\right\|_\infty \leq W_{\mathsf{max}}$. With probability at least $1 - \delta$, the maximum likelihood estimator $\widehat{\pi} := \arg\max_{\pi\in\Pi} \widehat{L}_n(\pi)$ satisfies,*

$$D_{\mathsf{KL}}(\pi_\mathsf{D} \,\|\, \widehat{\pi}) \lesssim \log W_{\mathsf{max}} \cdot \inf_{\varepsilon>0}\left\{\frac{\log \mathcal{N}_\infty(\Pi, \varepsilon)}{n} + \varepsilon\right\}, \tag{29}$$

and consequently

$$\mathsf{Cov}_M(\widehat{\pi}) \lesssim \frac{\log W_{\mathsf{max}}}{\log M} \cdot \inf_{\varepsilon>0}\left\{\frac{\log \mathcal{N}_\infty(\Pi, \varepsilon)}{n} + \varepsilon\right\}, \tag{30}$$

for all $M \geq 2$.

We remark that the $\log(W_{\mathsf{max}})$-factor in Eq. (29) can be tight in general. For example, for the class Π considered in Proposition 3.2, it holds that $\log \mathcal{N}_\infty(\Pi, \varepsilon) \lesssim \log(1/\varepsilon) \vee 1$ and $\left\|\frac{\pi_{\mathsf{D}}}{\pi}\right\|_\infty \leq e^{2H}$.

Proof of Proposition D.9. By Lemma 4 of Yang & Barron (1998), it holds that

$$D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) \leq (2 + \log(W_{\mathsf{max}})) D^2_{\mathsf{H}}(\pi_{\mathsf{D}}, \widehat{\pi}).$$

Therefore, the first bound then follows from Eq. (27). The second bound follows from applying Proposition D.1. □

D.4 Autoregressive Models: Coverage and Stopped KL-Divergence

Proposition D.10. *Define*

$$D_{\mathsf{seq},N}(\pi_{\mathsf{D}} \,\|\, \pi) = \mathbb{E}_{(x,y_{1:H})\sim\pi_{\mathsf{D}}} \min\left\{\log N, \sum_{h=1}^{H} D_{\mathsf{KL}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}) \,\|\, \pi(\cdot \mid x, y_{1:h-1}))\right\}. \quad (31)$$

Then for $N > e$, it holds that

$$\mathsf{Cov}_N(\pi_{\mathsf{D}} \,\|\, \pi) \leq \frac{2}{\log N - 1} D_{\mathsf{seq},N}(\pi_{\mathsf{D}} \,\|\, \pi). \quad (32)$$

Proof of Proposition D.10. Consider the stopping time

$$\tau := \min\left\{h : h = H \text{ or } \sum_{j\leq h} D_{\mathsf{KL}}(\pi_{\mathsf{D}}(y_{j+1} = \cdot \mid x, y_{1:j}) \,\|\, \pi(y_{j+1} = \cdot \mid x, y_{1:j})) > \log N\right\}.$$

Then, for the process $Y^\tau = (x, y_{1:\tau})$, we have the chain rule:

$$\begin{aligned}
&D_{\mathsf{KL}}(\pi_{\mathsf{D}}(Y^\tau = \cdot) \,\|\, \pi(Y^\tau = \cdot)) \\
&= \mathbb{E}_{\pi_{\mathsf{D}}}\left[\sum_{h=1}^{\tau} D_{\mathsf{KL}}(\pi_{\mathsf{D}}(y_h = \cdot \mid x, y_{1:h-1}) \,\|\, \pi(y_h = \cdot \mid x, y_{1:h-1}))\right] \\
&\leq \mathbb{E}_{\pi_{\mathsf{D}}} \min\left\{\log N, \sum_{h=1}^{H} D_{\mathsf{KL}}(\pi_{\mathsf{D}}(y_h = \cdot \mid x, y_{1:h-1}) \,\|\, \pi(y_h = \cdot \mid x, y_{1:h-1}))\right\},
\end{aligned}$$

where the inequality uses $\sum_{j<\tau} D_{\mathsf{KL}}(\pi_{\mathsf{D}}(y_{j+1} = \cdot \mid x, y_{1:j}) \,\|\, \pi(y_{j+1} = \cdot \mid x, y_{1:j})) \leq \log N$, which follows from the definition of τ. Therefore, by Proposition D.1, we have

$$\mathbb{P}_{\pi_{\mathsf{D}}}\left(\frac{\pi_{\mathsf{D}}(Y^\tau)}{\pi(Y^\tau)} \geq \log N\right) \leq \frac{D_{\mathsf{KL}}(\pi_{\mathsf{D}}(Y^\tau = \cdot) \,\|\, \pi(Y^\tau = \cdot))}{\log N - 1 + 1/N}.$$

Finally, we bound

$$\mathbb{P}_{\pi_{\mathsf{D}}}\left(\frac{\pi_{\mathsf{D}}(y_{1:H} \mid x)}{\pi(y_{1:H} \mid x)} \geq N\right) \leq \mathbb{P}_{\pi_{\mathsf{D}}}(\tau < H) + \mathbb{P}_{\pi_{\mathsf{D}}}\left(\frac{\pi_{\mathsf{D}}(Y^\tau)}{\pi(Y^\tau)} \geq \log N\right).$$

By Markov's inequality,

$$\begin{aligned}
\mathbb{P}_{\pi_{\mathsf{D}}}(\tau < H) &\leq \mathbb{P}_{\pi_{\mathsf{D}}}\left(\sum_{h=1}^{H} D_{\mathsf{KL}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}) \,\|\, \pi(\cdot \mid x, y_{1:h-1})) > \log N\right) \\
&\leq \frac{1}{\log N} \mathbb{E}_{\pi_{\mathsf{D}}} \min\left\{\log N, \sum_{h=1}^{H} D_{\mathsf{KL}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}) \,\|\, \pi(\cdot \mid x, y_{1:h-1}))\right\}.
\end{aligned}$$

Combining the inequalities above completes the proof. □

Proposition D.11. *For any $N \geq 1, \delta \in (0,1)$, it holds that*

$$\mathsf{Cov}_N(\pi_{\mathsf{D}} \,\|\, \pi) \geq \mathbb{P}_{\pi_{\mathsf{D}}}\left(\sum_{h=1}^{H} D^2_{\mathsf{H}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}), \pi(\cdot \mid x, y_{1:h-1})) \geq \log(N/\delta)\right) - \delta.$$

Proof of Proposition D.11. By definition,

$$\begin{aligned}
&\mathbb{E}_{y_h \sim \pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1})} \exp\left(-\frac{1}{2} \log \frac{\pi_{\mathsf{D}}(y_h \mid x, y_{1:h-1})}{\pi(y \mid x, y_{1:h-1})}\right) \\
&= \sum_{y_h \in \mathcal{Y}} \sqrt{\pi_{\mathsf{D}}(y_h \mid x, y_{1:h-1}) \cdot \pi(y \mid x, y_{1:h-1})} \\
&= 1 - D^2_{\mathsf{H}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}), \pi(\cdot \mid x, y_{1:h-1})) \leq \exp\big(-D^2_{\mathsf{H}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}), \pi(\cdot \mid x, y_{1:h-1}))\big).
\end{aligned}$$

Therefore, it holds that

$$\mathbb{E}_{\pi_{\mathsf{D}}} \exp\left(\sum_{h=1}^{H} D^2_{\mathsf{H}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}), \pi(\cdot \mid x, y_{1:h-1})) - \frac{1}{2} \log \frac{\pi_{\mathsf{D}}(y_h \mid x, y_{1:h-1})}{\pi(y \mid x, y_{1:h-1})}\right) \leq 1.$$

By Markov inequality, this implies

$$\mathbb{P}_{\pi_{\mathsf{D}}}\left(\frac{1}{2} \log \frac{\pi_{\mathsf{D}}(y_{1:H} \mid x)}{\pi(y_{1:H} \mid x)} \leq \sum_{h=1}^{H} D^2_{\mathsf{H}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}), \pi(\cdot \mid x, y_{1:h-1})) - \log(1/\delta)\right) \leq \delta.$$

To conclude, we note that

$$\begin{aligned}
&\mathbb{P}_{\pi_{\mathsf{D}}}\left(\sum_{h=1}^{H} D^2_{\mathsf{H}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}), \pi(\cdot \mid x, y_{1:h-1})) \geq \log(N/\delta)\right) \\
&\leq \mathbb{P}_{\pi_{\mathsf{D}}}\left(\sum_{h=1}^{H} D^2_{\mathsf{H}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}), \pi(\cdot \mid x, y_{1:h-1})) \geq \frac{1}{2} \log \frac{\pi_{\mathsf{D}}(y_{1:H} \mid x)}{\pi(y_{1:H} \mid x)} + \log(1/\delta)\right) \\
&\quad + \mathbb{P}_{\pi_{\mathsf{D}}}\left(\frac{1}{2} \log \frac{\pi_{\mathsf{D}}(y_{1:H} \mid x)}{\pi(y_{1:H} \mid x)} + \log(1/\delta) \geq \log(N/\delta)\right) \\
&\leq \delta + \mathsf{Cov}_N(\pi_{\mathsf{D}} \,\|\, \pi).
\end{aligned}$$

Re-organizing gives the desired result. □

E Comparison to Classical Generalization Bounds for Maximum Likelihood

In this section we briefly compare our main coverage-based generalization bound for maximum likelihood to classical generalization bounds for maximum likelihood based on Hellinger distance and KL-divergence.

Comparison to KL concentration. For general model classes Π, the best KL generalization bound we are aware of is Proposition D.9 (Appendix D), which scales as roughly

$$D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) \lesssim \log W_{\mathsf{max}} \cdot \mathcal{C}_{\mathsf{fine}}(\Pi, n)$$

under the assumption that all $\pi \in \Pi$ obey a sequence-level density ratio bound $\left\|\frac{\pi_{\mathsf{D}}}{\pi}\right\|_\infty \leq W_{\mathsf{max}}$, where $\log \mathcal{N}$ is an apropriate notion of covering number; note that for the autoregressive linear class, we have $\log W_{\mathsf{max}} = BH$, matching Proposition 3.2. Combining such a guarantee with Proposition 3.1 gives a coverage bound of roughly

$$\mathsf{Cov}_N(\widehat{\pi}) \lesssim \frac{\log W_{\mathsf{max}}}{\log N} \cdot \mathcal{C}_{\mathsf{fine}}(\Pi, n);$$

this is rather uninteresting since $\mathsf{Cov}_N(\widehat{\pi}) = 0$ for $N \geq W_{\mathsf{max}}$; in other words, we do not get a meaningful improvement as we scale N.

Asymptotic bounds for maximum likelihood. We also note that the classical theory of maximum likelihood (e.g., Van der Vaart (2000)) provides the following *asymptotic* convergence rate for d-dimensional parametric classes Π:

$$D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) \lesssim \frac{d}{n} \lesssim \mathcal{C}_{\mathsf{fine}}(\Pi, n), \qquad n \to +\infty.$$

While this upper bound does *not* scale with $\log W_{\mathsf{max}}$, it can only be attained with $n \geq n_0$ for a sufficiently large burn-in cost n_0 (typically scaling with $\log W_{\mathsf{max}}$ itself or similar problem-dependent parameters; see, e.g., Spokoiny (2012) for non-asymptotic bounds of this type).

Comparison to Hellinger concentration. For general model classes Π, the best Hellinger generalization bound we are aware of is Proposition D.8 (Appendix D), which scales as roughly

$$D^2_{\mathsf{H}}(\pi_{\mathtt{D}}, \widehat{\pi}) \lesssim \mathcal{C}_{\mathtt{fine}}(\Pi, n)$$

Combining such a guarantee with Proposition 3.1 gives a coverage bound of roughly

$$\mathsf{Cov}_N(\widehat{\pi}) \lesssim \mathcal{C}_{\mathtt{fine}}(\Pi, n)$$

for all $N \geq 2$. Compare to the KL-based result above, this result gives a non-trivial bound on coverage when N is constant (comparable to Theorem 4.1), but the issue is that it gives no further improvement as we scale N.

F Additional Results

F.1 Tightness of Theorem 4.1

To conclude, we show that the coarse and fine-grained terms in Theorem 4.1 are both tight in general.

Proposition F.1. *The following lower bounds on coverage hold for the maximum likelihood estimator.*

(a) Coarse rate: *For any $n, d \geq 1$ and $B \geq \log(5n)$, there exists a class Π with $\log \mathcal{N}_\infty(\Pi, \alpha) \lesssim d\log(B/\alpha) \vee 1$ and $\pi_{\mathsf{D}} \in \Pi$ such that with probability at least 0.5, it holds that for any $N \leq e^B$,*

$$\mathsf{Cov}_N(\widehat{\pi}) \geq c \cdot \frac{d}{n}.$$

(b) Fine rate: *For any $n \geq d \geq 1, N \geq 1$, there exists a class Π and $\pi_{\mathsf{D}} \in \Pi$ such that $|\Pi| = 2^d + 1$ and $\mathcal{N}_\infty(\Pi, \alpha) \leq 2$ for any $\alpha \geq \sqrt{\frac{d}{n}}$, and with probability at least 0.5, it holds that*

$$\mathsf{Cov}_N(\widehat{\pi}) \geq c \cdot \frac{d}{n \cdot \log N}.$$

Informally, case (a) shows that for the class Π under consideration, the coverage does not decrease with $\log N$ until N is trivially large such that $\log \mathcal{N}_\infty(\Pi, \log N) = 0$; this is precisely the behavior of the coarse term in Theorem 4.1, so this implies there is no hope of removing this term. Meanwhile, case (b) can be interpreted as showing that there is no hope of replacing the high-precision covering number found in the fine-grained term in Theorem 4.1 with a coarser notion (e.g, at the scale in the coarse-grained term), since the rate grows with $d \approx \log|\Pi|$ even though $\log \mathcal{N}(\Pi, \alpha)$ is constant for $\alpha \geq \sqrt{\frac{d}{n}}$. We note that Proposition F.1 is an algorithm-specific lower bound, not an information-theoretic lower bound; we show in Section 6.2 *is* that it is possible to improve over Theorem 4.1 with algorithms explicitly designed to optimize for coverage.

F.2 Maximum Likelihood: Tighter Rates for Convex Classes

In the following, we analyze the MLE for *convex* model class.

Assumption F.1 (Convex model class)**.** *The class Π satisfies $\Pi = \{\pi_\theta : \theta \in \Theta\}$ for a convex, compact parameter space Θ, and the mapping $\theta \mapsto \pi_\theta(y \mid x)$ is concave for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$.*

Theorem F.1 (Fast convergence of coverage for convex classes)**.** *Let $\alpha \geq 0$, $N' \geq 1$, $N \geq 2e^{2\alpha}N'$ be given, and suppose that Assumption F.1 holds. Define*

$$\theta^\star = \arg\min_{\theta \in \Theta} D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \pi_\theta).$$

With probability at least $1-\delta$, the maximum likelihood estimator $\widehat{\pi} := \arg\max_{\pi \in \Pi} \widehat{L}_n(\pi)$ satisfies

$$\mathsf{Cov}_N(\widehat{\pi}) \leq \mathsf{Cov}_{N'}(\pi_{\theta^\star}) + C\frac{\log \mathcal{N}_\infty\left(\Pi; \alpha\right) + \log(\delta^{-1})}{n} + \frac{Ce^{2\alpha}N'}{N} \cdot \inf_{\varepsilon > 0}\left\{\frac{\log \mathcal{N}_\infty(\Pi, \varepsilon)}{n} + \varepsilon\right\}, \tag{33}$$

where $C > 0$ is an absolute constant.

In words, we show that for convex class, the coverage of MLE $\widehat{\pi}$ can be upper bounded by the coverage of $\pi_{\theta^\star}$, the best-in-class approximate of π_{D}. In particular, when $\pi_{\mathsf{D}} \in \Pi$, we get the following bound for convex class Π:

$$\begin{aligned}\mathsf{Cov}_N(\widehat{\pi}) &\lesssim \frac{1}{N} \cdot \inf_{\varepsilon > 0}\left\{\frac{\log \mathcal{N}_\infty(\Pi, \varepsilon)}{n} + \varepsilon\right\} + \frac{\log \mathcal{N}_\infty\left(\Pi, c\log N\right) + \log(\delta^{-1})}{n} \\ &= \frac{\mathcal{C}_{\mathsf{fine}}(\Pi, n)}{N^{1-2c}} + \mathcal{C}_{\mathsf{coarse}}(\Pi, N, n),\end{aligned}$$

which improves upon the bound $\mathsf{Cov}_N(\widehat{\pi}) \lesssim \frac{\mathcal{C}_{\mathsf{fine}}(\Pi, n)}{\log N} + \mathcal{C}_{\mathsf{coarse}}(\Pi, N, n)$ shown in Theorem 4.1 for general class Π. The proof is presented in Appendix J.3.

F.3 Stochastic Gradient Descent: Improved Gradient Normalization for Distillation

In this section, we focus on autoregressive linear models (2), and consider a variant of our setting inspired by distillation . We assume that for each example $(x^i, y^i_{1:H})$, for each $h = 1, \ldots, H$, we have access to the true next-token probabilities $\pi_{\mathsf{D}}(y_h \mid x^i, y^i_{1:h-1})$ for all $y_h \in \mathcal{V}$. This is an unrealistic assumption for general pre-training, but it is natural for distillation, where π_{D} corresponds to a teacher model (in particular, the next-token probabilities are already computed as part of a standard forward pass through the teacher model).

For the distillation setting, we give an improved gradient normalization scheme that improves upon the rate achieved by Theorem 5.1, closing the gap between SGD and maximum likelihood by matching the guarantee for Theorem 4.2.

Define $\epsilon_\theta(x, y_{1:h-1}) := D_{\mathsf{KL}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}) \,\|\, \pi_\theta(\cdot \mid x, y_{1:h-1}))$; note that for the distillation setting, we can compute this quantity in closed form for any prefix $x, y_{1:h-1}$ in the training corpus. We consider the following (single-sample) truncated/normalized stochastic gradient estimator:

$$\widehat{g}_\theta(y \mid x) = \sum_{h=1}^{H} \alpha_\theta(x, y_{1:h-1}) \nabla \log \pi_\theta(y_h \mid x, y_{1:h-1}), \tag{34}$$

where $A := \log N$, and where

$$\alpha_\theta(x, y_{1:h-1}) = \begin{cases} 1, & \sum_{j \le h-1} \epsilon_\theta(x, y_{1:j}) \le A, \\ 0, & \sum_{j < h-1} \epsilon_\theta(x, y_{1:j}) > A, \\ \frac{A - \sum_{j<h-1} \epsilon_\theta(x, y_{1:j})}{\epsilon_\theta(x, y_{1:h-1})}, & \text{otherwise.} \end{cases} \tag{35}$$

With this definition, we define the following normalized SGD update:

$$\theta^{t+1} = \mathrm{Proj}_\Theta(\theta^t + \eta \widehat{g}_{\theta^t}(y^t \mid x^t)). \tag{36}$$

Intuitively, the idea behind the update in Eq. (34) is to truncate the gradient at the point where the KL divergence between the teacher and student model is too large, and then normalize the gradient by the KL divergence; this is inspired by the structural result Proposition D.10 in Appendix D.4, where we show a close connection between the coverage profile and a certain "stopped" variant of KL divergence.

Theorem F.2. *Let $T, N \ge 1$ be given. With a suitably chosen stepsize $\eta > 0$, the normalized SGD update (9) achieves the following coverage bound:*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathsf{Cov}_N(\pi_{\theta^t})\right] \lesssim \sqrt{\frac{\sigma_\star^2}{T \log N}} + \frac{B^2}{T}. \tag{37}$$

This guarantee matches the rate of Theorem 4.2 for the maximum likelihood estimator. The proof is presented in Appendix L.2.

F.4 An improved tournament via on-policy generation

We describe an improved tournament estimator that is able to remove that $1/N^{1-a}$ term from Theorem 6.2, meaning it achieves nontrivial guarantees even when the coverage parameter N is a constant.

Note that the term $1/N^{1-a}$ of Eq. (14) comes from the fact that $\mathbb{P}_{\pi_{\mathsf{D}}}(\frac{\pi(y|x)}{\pi_{\mathsf{D}}(y|x)} \ge N)$ can be as large as $1/N$ in the worst case, implying that the $\widehat{\pi}$ produced by Eq. (13) may at best achieve a coverage of $1/N$. To overcome this, we introduce an *offset term*:

$$\widehat{\pi} := \arg\min_{\pi \in \Pi} \max_{\pi' \in \Pi} \left\{ \widehat{\mathsf{Cov}}_N(\pi' \,\|\, \pi) - 2N^a \cdot \widehat{\mathsf{Cov}}^{\overline{\pi}}_N(\pi' \,\|\, \pi) \right\}, \tag{38}$$

where we define $\widehat{\mathsf{Cov}}^{\overline{\pi}}_N(\pi' \,\|\, \pi) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_{y \sim \overline{\pi}(\cdot|x^i)}\left(\frac{\pi'(y|x^i)}{\pi(y|x^i)} \ge N\right)$ for models $\pi, \pi', \overline{\pi}$. This estimator augments the simple tournament in Eq. (13) with an "offset" term that accounts for the fact that some of the models might be quite far from π_{D}. The main guarantee is as follows.

Theorem F.3. *Fix $N \geq 1$, $a > 0$ such that $N^{1-2a} \geq 4$. Suppose that there exists $\overline{\pi} \in \Pi$ such that $|\log \pi_{\mathsf{D}}(y \mid x) - \log \overline{\pi}(y \mid x)| \leq a \log N$ for any $x \in \mathcal{X}, y \in \mathcal{Y}$. Then with probability $1 - \delta$, the tournament estimator (38) achieves $\mathsf{Cov}_{2N^{1+a}}(\widehat{\pi}) \lesssim \frac{\log(|\Pi|/\delta)}{n}$.*

Compared to Theorem 6.2, this tournament eliminates the additive $1/N^{1-a}$ term. It does, however, require a stronger condition on the best-in-class model $\overline{\pi}$ that $|\log \pi_{\mathsf{D}}(y \mid x) - \log \overline{\pi}(y \mid x)| \leq a \log N$, which implies in particular that $\mathsf{Cov}_{N^a}(\overline{\pi}) = 0$.

Infinite classes: Beating maximum likelihood. While we motivated the tournament estimators through model/checkpoint selection with a finite class Π, both estimators can also be applied to general, infinite classes Π. In this case, it turns out that they both improve upon the coverage achieved by the maximum likelihood estimator in Theorem 4.1, even in the well-specified case where $\pi_{\mathsf{D}} \in \Pi$; informally, the tournament estimators allow us to remove the fine-grained term in Theorem 4.1, leaving only a coarse-grained term. See Theorem 6.2′ and Theorem F.3′ for the formal statements.

G EXPERIMENTS

We describe the general graph search task used throughout our experiments in Appendix G.1, then detail the specific setups used for Figure 1 in Appendix G.2, and for Figure 2 in Appendix G.3.

G.1 GRAPH REASONING TASK

We evaluate our theoretical predictions using experiments in graph reasoning tasks, in which transformer models are trained to find paths between source and target nodes in graphs. Both graph reasoning benchmarks and synthetic datasets have seen increasing use as abstractions for reasoning problems and for probing language modeling phenomena (Sanford et al., 2024; Nagarajan et al., 2025; Saparov et al., 2025; Bachmann & Nagarajan, 2024; Yehudai et al., 2025; Taylor et al., 2024; Wang et al., 2023; Fatemi et al., 2024; Tang et al., 2025).

These tasks provide minimal abstractions of core reasoning problems, yet are expressive enough to capture pre-training and fine-tuning phenomena. They also offer flexibility in problem structure and difficulty: by specifying different graph topologies and path depths, we can modulate difficulty and expose sources of hardness. At the same time, the simplicity of the setup enables training in controlled settings with interpretable results with which to ground our theoretical predictions.

G.1.1 GRAPH SEARCH TASK DESCRIPTION

The graph search tasks in Appendix G.2 and Appendix G.3 share the same high-level structure, and are comprised of

- A set of graph structures $\mathcal{G}$ that map bijectively to a set of prompts $\mathcal{X}$, and induce the response space $\mathcal{Y}$
- A distribution over the prompts $\mu \in \Delta(\mathcal{X})$
- A data collection policy $\pi_{\mathsf{D}} : \mathcal{X} \to \Delta(\mathcal{Y})$

Next, we describe the general details of the graph search task common to all experiments, and leave the task details for each figure in the proceeding sections.

Graph Search. The nodes of all graphs in a given task $\mathcal{G}$ are drawn from the set $[m]$, for some integer $m \in \mathbb{N}$. Each graph structure $G = (V, E) \in \mathcal{G}$ is comprised of a set of vertices ("nodes") $V \subseteq [m]$ and edges $E = \{(u, v) : u, v \in V, u \neq v\}$. It also contains one source node $s \in V$ and one target node $t \in V$, so that (G, s, t) specifies one search problem instance within the class. The search task can be translated into an autoregressive sequence modeling problem using the below prompt and response specifications.

Layered Graph Structure. For all experiments, we utilize a *layered directed acyclic graph (layered DAG)* with a rectangular structure. Following the source node, the graph has L layers each with a fixed number of nodes. In each layer, only a subset of its nodes has edges connecting to the next layer, and we refer to these nodes as *passable nodes*, or the set $\left\{v \in V : \deg^+(v) > 0\right\}$ that has non-zero out-degree. In a given layer each passable node has edges to all nodes in the next layer, while the remaining (non-passable) nodes in the layer cannot be used to traverse to the target, so that as soon as the model outputs one such node its path cannot be valid.

In order to output a valid path from source to target, it is sufficient to keep outputting the passable nodes in the next layer. In general a graph may have many valid paths, and in each experiment, π_{D} always samples valid paths. However, as will describe shortly, π_{D} may use complex functions to sample from only a subset of the valid paths, and $\widehat{\pi}$ must learn to cover such behavior.

The layered DAG offers a natural interpretation as an abstraction for reasoning problems. Following passable nodes in valid paths corresponds to taking reasoning steps that make progress towards the solution, and selecting paths via more complex functions maps to learning high-quality solutions that accurately reflect desired properties for the problem's solution.

Graphs to Prompts. Given a graph structure $G = (V, E) \in \mathcal{G}$ and a source node $s \in V$ and target node $t \in V$, the prompt x encodes the search problem as the adjacency list of G with the source and target nodes appended to the end. The prompt is formatted as a string of the form

$$x: \texttt{ u\_1 v\_1 | u\_2 v\_2 | . . . | u\_k v\_k / s t =}$$

where $(u_i, v_i) \in [m]^2$ are the vertices of the i-th edge in the adjacency list. For example, with edges $E = \{(10, 23), (86, 47), \ldots, (45, 32)\}$, the prompt is formatted as

$$x: \texttt{ 10 23 | 86 47 | . . . | 45 32 / 10 45 =}$$

where | and / are special characters that separate two edges and the adjacency list from the source and target nodes, respectively. The special character = is indicates the end of the prompt.

Graphs to Responses. Given a graph structure $G = (V, E) \in \mathcal{G}$ and a source node $s \in V$ and target node $t \in V$, the response y encodes the path from the source to the target node in G. In general a graph may have multiple paths from source to target. The horizon H corresponds to the longest path length in $\mathcal{G}$, and a response takes the form of a string

$$y: \texttt{ s v\_1 v\_2 v\_3 . . . v\_H t}$$

where $v_i \in [m]$ are the vertices of the i-th edge in the path.

Sequence Modeling Problem. In summary, a graph search task with set of graphs $\mathcal{G}$ induces an autoregressive sequence modeling problem with a vocabulary space $\mathcal{V} = [m] \cup \{|, /, =\}$, prompts $\mathcal{X} \subseteq \mathcal{V}^*$ corresponding to graph structures, and responses $\mathcal{Y} \subseteq \mathcal{V}^H$ corresponding to paths with length at most H. In addition, the task is equipped with $\mu \in \Delta(\mathcal{X})$ and $\pi_{\mathtt{D}} : \mathcal{X} \to \Delta(\mathcal{Y})$ that is used to collect the training dataset $\mathtt{D} = \{(x, y)\}$, where $x \sim \mu$ and $y \sim \pi_{\mathtt{D}}(x)$.

G.1.2 General Implementation Details

Next, we describe the common implementation details of the graph search task.

Tokenizer. The tokenizer is a numeral tokenizer standard for graph reasoning tasks. Each node $v \in [m]$ is tokenized as its integer node value, and the special characters |, /, and = are tokenized as $m + 1, m + 2, m + 3$, respectively.

Transformer model. Throughout our experiments, we train causally-masked GPT2 transformer models to minimize the cross-entropy loss using the Adam optimizer with fixed learning rate, and perform a grid search over the parameters displayed in Table 1. Parameters with fixed values were chosen based on related papers such as Bachmann & Nagarajan (2024). In both experiments, the model architecture with 4 heads, 6 hidden layers, and 384 hidden dimensions worked best. We use absolute positional encodings. Training iterations and grid search values for the learning rate are different for each experiment, and discussed further below.

Hyperparameter	Values
Number of heads	{4, 6, 8}
Number of layers	{3, 4, 6, 8}
Hidden dimensions	384
Activation function	GeLU
Batch size	128
Weight decay	0.01

Table 1: Hyperparameter grid search values for transformer models in graph search.

G.2 Experiment Details for Figure 1

The graph search task for Figure 1 exposes natural properties of pre-training data where cross-entropy reduction comes at the cost of a worse coverage profile. In particular, because pre-training data is diverse, the model in practice is generally unable to perfectly fit the distribution. When one mode of behavior is better-represented than another, cross-entropy minimization, which is an average-case distribution-matching metric, can sacrifice coverage over the different modes in order to increase performance on one.

Correspondingly, our graph search task for Figure 1 is a mixture of two classes of graph structures. Due to representational and finite-sample constraints, the model is unable to fit both perfectly during training, and, in particular, fitting one class well (in the sense of cross-entropy loss) comes at the cost of worse performance on the other. The checkpoint with the best coverage arises at some middle point in training when the model learns both classes of graphs equally well, and has good coverage over

both classes (the dip $\texttt{Cov}_N$ in the leftmost subplot of Figure 1). Further reduction of cross-entropy loss over the latter half of training requires the model to lose coverage over $\pi_{\texttt{D}}$ in the less-represented graph class (observed as the increase in $\texttt{Cov}_N$ in the latter half of training iterations).

Even though the task cannot be learned perfectly from the supervised learning feedback, the model can still learn a policy that always samples a correct path matching $\pi_{\texttt{D}}$'s with $O(1)$ sampling attempts, which means that it leads to efficient downstream post-training (e.g., on one of modes or with reward-based feedback), and also achieve optimal performance with test-time inference methods.

For the experiments in Figure 1, we first pre-train a model on a larger set of graph structure classes so that it learns a diverse set of behaviors, then finetune its behavior on two. The performance on the finetuning task is displayed in Figure 1, and we first describe the finetuning dataset, followed by the pre-training dataset.

G.2.1 Task Description

The finetuning task is a skewed mixture over two graph disjoint graph classes, $\mathcal{G}_1 \cup \mathcal{G}_2 = \mathcal{G}$, with $\widetilde{\mu} \in \Delta(\{1, 2\})$ denoting the probability of each class in the data. All graphs in $\mathcal{G}$ follow the *layered DAG* structure described in Appendix G.1, with $L = 8$ layers that each have 4 nodes. Of the 8 layers, 2 are randomly selected to have 2 passable nodes (meaning that they are connected to the next layer), while the remaining layers have only 1 passable node. Although there are 4 valid paths from source to target, the policy $\pi_{\texttt{D}}$ is deterministic and chooses 1 based on a rule, which is what distinguishes the two class types described below.

Class $\mathcal{G}_1$ with probability $\widetilde{\mu}(1) = 0.9$. For an integer $j \in \mathbb{Z}$, let the function $p(j) = (j \mod 2)$ denote its parity. In layers with 1 passable node, $\pi_{\texttt{D}}$ takes the passable node. For each layer $l \in [L]$ with 2 passable nodes (there is guaranteed to be one even and one odd), $\pi_{\texttt{D}}$ chooses the node v such that $p(v) = p(l)$, that is, the node whose parity is equal to the parity of the layer index.

Class $\mathcal{G}_2$ with probability $\widetilde{\mu}(2) = 0.1$. In this class $\pi_{\texttt{D}}$ takes the opposite rule from the one in $\mathcal{G}_2$: for layers l with 2 passable nodes, it chooses the node v such that $p(v) = 1 \oplus p(l)$.

The class of a graph is technically identifiable from the prompt, but the problem is too difficult for the model to learn in just the finetuning stge. The class of a graph can be computed from the parity of a hidden subset of their nodes whose cardinality is half the total number of nodes; letting this hidden subset be $V' \in V$, all graphs in $\mathcal{G}_1$ have $1 = \bigoplus_{u \in V'} p(u)$, while the opposite is true for all graphs in $\mathcal{G}_2$.

Dataset generation. Each sample in the dataset $D = \{(x, y)\}$ is then generated via the following procedure.

1. First sample an index $i \sim \widetilde{\mu}$.
2. Sample $G \in \mathcal{G}_i$ by randomly drawing $V \subset [m]$ without replacement, and instantiate the edges according to the description for each class above.
3. Format the prompt x per Appendix G.1.
4. Draw $y \sim \pi_{\texttt{D}}(\cdot \mid x)$ according to description for each class above.

G.2.2 Pre-training Description

The graphs in the pre-training task are a superset of the graphs in the finetuning task, that is, $\cup_{i \in [K]} \mathcal{G}_i = \mathcal{G}$ with $K = 3$, and the data distribution is a uniform mixture of these 3 classes, $\widetilde{\mu}(i) = \frac{1}{K}$ for each $i \in [K]$. The first two classes $\mathcal{G}_1$ and $\mathcal{G}_2$ are defined exactly as they are in the finetuning dataset. In $\mathcal{G}_3$, two layers have 2 passable nodes, while the rest have 1, and $\pi_{\texttt{D}}$ samples one of the 2^2 valid paths from source to target at random. The dataset is sampled using the same dataset generation procedure described for the finetuning task above.

G.2.3 Task-specific Implementation Details

The transformer model is first pre-trained on a fixed dataset drawn from the pre-training distribution, with $8 \times 64,000$ prompts in total, using a learning rate of $1\text{e}{-4}$ for 200k iterations, which was chosen based on a grid search over learning rates $\{5\text{e}{-5}, 1\text{e}{-4}, 5\text{e}{-4}\}$.

The final checkpoint is then finetuned for 50k iterations in an online fashion, where fresh samples are drawn for each batch (this is equivalent to offline training with a dataset tha has an equivalent number

of samples). The learning rate is $5\text{e}{-6}$, which was chosen based on a grid search over learning rates $\{5\text{e}{-6}, 1\text{e}{-5}\}$.

G.3 Experiment Details for Figure 2

The graph structures used for Figure 2 to expose the dependence on horizon of KL-divergence but not Cov_N leverages the intuition from Remark 3.1. The training data is homogeneous, and a fraction consists of difficult graph problems that the learner cannot cover with the given finite samples. The coverage profile Cov_N will be the same regardless of H, but KL scales linearly with H due to this unlearnable subset of the data.

G.3.1 Task Description

For Figure 2, we devise a family of tasks that is defined per H, which we then instantiate with $H \in \{8, 16, 24\}$ for our results. For a fixed H, the task $\mathcal{G}_H$ utilizes the *layered DAG* graph structure described in Appendix G.1 with H layers of 4 nodes each, so that $\mathcal{Y} = \mathcal{V}^{H+2}$ when the source and target nodes are included.

The task is a heterogeneous mixture over 3 classes of graphs described below that we refer to as $\mathcal{G}_{H,1} \cup \mathcal{G}_{H,2} \cup \mathcal{G}_{H,3} = \mathcal{G}_H$. The classes $\mathcal{G}_{H,2}$ and $\mathcal{G}_{H,3}$ are significantly harder to learn and the model will fail to do so with the given number of training samples, even though $\mathcal{G}_{H,1}$ is learned quickly (and also provides useful features for learning the other two tasks). The distribution over these 3 classes is fixed for all H and specified by $\widetilde{\mu} \in \Delta(\{1, 2, 3\})$.

Class $\mathcal{G}_{H,1}$ with probability $\widetilde{\mu}(1) = 0.94$. All H layers have only 1 passable node, so each $G \in \mathcal{G}_{H,1}$ has only one valid path from source to target. For prompts corresponding to graphs in this class, π_{D} deterministically takes the single valid path.

Class $\mathcal{G}_{H,2}$ with probability $\widetilde{\mu}(2) = 0.05$. Half of the layers (or $H/2$, selected randomly) have 2 passable nodes while the rest have 1. While there are $2^{H/2}$ valid paths from source to target, π_{D} deterministically selects one of them. For layers with 2 passable nodes, one is guaranteed to be even and the other odd. In layers with more than one passable node, π_{D} selects one node by following a difficult, deterministic rule. This rule requires π_{D} to select the node v whose parity matches the parity of the layer index, XOR'ed with the parity of each passable node in the entire graph. More specifically, recall that $p(j)$ denotes the parity of an integer $j \in [m]$, and let $V^\star := \{v \in V : \deg^+(v) > 0\} \subset V$ denote the set of all passable nodes (or those with positive out-degree). Then in layer l, π_{D} selects the node v such that $p(v) = p(l) \oplus \left(\bigoplus_{u \in V^\star} p(u)\right)$.

Class $\mathcal{G}_{H,3}$ with probability $\widetilde{\mu}(3) = 0.01$. Regardless of H, 4 of the layers are randomly chosen to have 2 passable nodes, so that there are $2^4 = 16$ valid paths from source to goal Here, however, π_{D} samples one of the $2^{\frac{H}{2}}$ valid paths uniformly at random.

Note that prompts/graphs from each class are distinguishable from each other (or, identifiable) based on prompt features alone, so a powerful-enough model can acheve perfect performance across all of them simultaneously. $\mathcal{G}_{H_2}$, for example, has more edges and thus a longer prompt than $\mathcal{G}_{H_1}$; similar statements apply to $\mathcal{G}_{H_3}$. Dataset generation occurs in the same manner as described in Appendix G.2.

G.3.2 Task-specific Implementation Details

Here, we describe experiment-specific implementation details on top of those previously described in Appendix G.1 (which apply to all figures).

In addition to a grid search over the parameters in Table 1, we perform a search over learning rates $\{5\text{e}{-5}, 1\text{e}{-4}, 5\text{e}{-4}\}$, for which the learning rate of $1\text{e}{-4}$ exhibited the best validation performance. The model is trained for 40k iterations over a fixed dataset of $8 \times 64,000$ samples.

The results in Figure 2 are computed from evaluations of training checkpoints on per-class validation datasets of 1024 prompts from each $\mathcal{G}_{H_i}$; these metrics are then averaged according to the probabilities in $\widetilde{\mu}$ to obtain the final result. In total we ran 16 seeds, and plot their median. The shaded region in Figure 2 displays the region between the $\frac{1}{16}$ quantile and $\frac{15}{16}$ quantile.

Part II

Proofs

H TECHNICAL TOOLS

H.1 CONCENTRATION INEQUALITIES

Lemma H.1 (Azuma-Hoeffding)**.** *Let $(Z^i)_{i\le n}$ be a sequence of real-valued random variables adapted to a filtration $(\mathscr{F}_i)_{i\le n}$. If $|Z^i| \le R$ almost surely, then with probability at least $1-\delta$, for all $n' \le n$,*

$$\left|\sum_{i=1}^{n'} Z^i - \mathbb{E}_{i-1}[Z^i]\right| \le R\cdot\sqrt{8n\log(2\delta^{-1})}.$$

Lemma H.2 (Freedman's inequality)**.** *Let $(Z^i)_{i\le n}$ be a real-valued martingale difference sequence adapted to a filtration $(\mathscr{F}_i)_{i\le n}$. If $|Z^i| \le R$ almost surely, then for any $\eta \in (0, 1/R)$, with probability at least $1-\delta$, for all $n' \le n$,*

$$\sum_{i=1}^{n'} Z^i \le \eta\sum_{i=1}^{n'}\mathbb{E}_{i-1}\left[(Z^i)^2\right] + \frac{\log(\delta^{-1})}{\eta}.$$

The next result is a standard consequence of Lemma H.2 (e.g., Foster et al. (2021)).

Lemma H.3. *Let $(Z^i)_{i\le n}$ be a sequence of random variables adapted to a filtration $(\mathscr{F}_i)_{i\le n}$. If $0 \le Z^i \le R$ almost surely, then with probability at least $1-\delta$, for all $n' \le n$,*

$$\sum_{i=1}^{n'} Z^i \le \frac{3}{2}\sum_{i=1}^{n'}\mathbb{E}_{i-1}[Z^i] + 4R\log(2\delta^{-1}), \tag{39}$$

and

$$\sum_{i=1}^{n'}\mathbb{E}_{i-1}[Z^i] \le 2\sum_{i=1}^{n'} Z^i + 8R\log(2\delta^{-1}). \tag{40}$$

Lemma H.4. *Suppose that μ is a distribution over $\mathcal{Z}$, function class $\mathcal{F} \subseteq (\mathcal{Z}\to\mathbb{R})$ is given. We let $N(\mathcal{F},\epsilon;\|\cdot\|_\infty)$ be the ϵ-covering number of $\mathcal{F}$ under the norm $\rho(f,f') := \sup_{z\in\mathcal{Z}}|f(z)-f'(z)|$. Then, under $\mathcal{D} = \{Z^1,\cdots,Z^n\}$ drawn from μ i.i.d, the following holds with probability at least $1-\delta$:*

$$\sum_{i=1}^{n} f(Z^i) \le n\log\mathbb{E}_\mu[\exp(f(Z))] + \inf_{\epsilon\ge 0}\{\log N(\mathcal{F},\epsilon;\|\cdot\|_\infty) + 2Ln\epsilon\}.$$

Lemma H.5. *For distribution $P, Q \in \Delta(\mathcal{X})$, function $f : \mathcal{X}\to[-B,B]$, it holds that*

$$|\mathbb{E}_P[f] - \mathbb{E}_Q[f]|^2 \le 3\mathrm{Var}_Q[f]\cdot D^2_{\mathsf{H}}(P,Q) + 8B^2 D_{\mathsf{H}}(P,Q)^4.$$

Therefore, for any $f : \mathcal{X}\to\mathbb{R}^d$ with $\|f\| \le B$, it holds that

$$\|\mathbb{E}_P[f] - \mathbb{E}_Q[f]\| \le 2\sqrt{\mathbb{E}_Q\|f-\mathbb{E}_Q[f]\|^2}\cdot D_{\mathsf{H}}(P,Q) + 3BD^2_{\mathsf{H}}(P,Q), \tag{41}$$

and

$$\mathbb{E}_P\|f-\mathbb{E}_P[f]\|^2 \le 3\,\mathbb{E}_Q\|f-\mathbb{E}_Q[f]\|^2 + 8B^2D^2_{\mathsf{H}}(P,Q). \tag{42}$$

Proof of Lemma H.5. We denote $P(x)$ ($Q(x)$) to be the density function of P (Q). Then

$$|\mathbb{E}_P[f] - \mathbb{E}_Q[f]|^2 = \left(\int_{\mathcal{X}} (f(x) - \mathbb{E}_Q[f])(P(x) - Q(x))dx\right)^2 \tag{43}$$

$$\leq \int_{\mathcal{X}} (f(x) - \mathbb{E}_Q[f])^2(\sqrt{P(x)} + \sqrt{Q(x)})^2 dx \cdot \int_{\mathcal{X}} (\sqrt{P(x)} - \sqrt{Q(x)})^2 dx \tag{44}$$

$$\leq 4D^2_{\mathsf{H}}(P, Q) \cdot \left(\mathrm{Var}_Q[f] + \mathbb{E}_P(f - \mathbb{E}_Q[f])^2\right) \tag{45}$$

Further, we know

$$\mathbb{E}_P(f - \mathbb{E}_Q[f])^2 \leq 3\,\mathbb{E}_Q(f - \mathbb{E}_Q[f])^2 + 8B^2 D^2_{\mathsf{H}}(P, Q). \tag{46}$$

This gives the desired upper bound. □

Lemma H.6. *Suppose that $\phi : \mathcal{Y} \to \mathbb{B}_2(B)$ with $B \geq 1$, and for any $\theta \in \mathbb{B}_2(1)$, $\pi_\theta \in \Delta(\mathcal{Y})$ is defined as $\pi_\theta(y) \propto \exp(\langle \phi(y), \theta \rangle)$. Then for any $\theta^\star, \theta \in \mathbb{B}_2(1)$, it holds that*

$$\mathbb{E}_{y\sim\pi_{\theta^\star}} \langle \phi(y) - \mathbb{E}_{\pi_{\theta^\star}}[\phi], \theta - \theta^\star \rangle^2 \leq 15B D_{\mathsf{KL}}(\pi_{\theta^\star} \,\|\, \pi_\theta).$$

Proof. Denote $\bar{\phi}(y) := \phi(y) - \mathbb{E}_{\pi_{\theta^\star}}[\phi]$. By definition,

$$D_{\mathsf{KL}}(\pi_{\theta^\star} \,\|\, \pi_\theta) = \log \mathbb{E}_{y\sim\pi_{\theta^\star}}\left[\exp\left(\langle \bar{\phi}(y), \theta - \theta^\star \rangle\right)\right] \geq B \log \mathbb{E}_{y\sim\pi_{\theta^\star}}\left[\exp\left(\frac{1}{B}\langle \bar{\phi}(y), \theta - \theta^\star \rangle\right)\right]$$

Note that for $x \geq -4$, we have $e^x \geq 1 + x + \frac{1}{10}x^2$. Therefore, we have

$$\frac{1}{B} D_{\mathsf{KL}}(\pi_{\theta^\star} \,\|\, \pi_\theta) \geq \log\left(1 + \frac{1}{10B^2}\,\mathbb{E}_{y\sim\pi_{\theta^\star}} \langle \bar{\phi}(y), \theta - \theta^\star \rangle^2\right) \geq \frac{1}{15B^2}\,\mathbb{E}_{y\sim\pi_{\theta^\star}} \langle \bar{\phi}(y), \theta - \theta^\star \rangle^2,$$

where we use $\log(1 + x) \geq \frac{3}{4}x$ for all $x \in [0, \frac{8}{5}]$. □

I Proofs from Section 3

Proof of Proposition 3.2. Consider the setting $d = 1$, $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{-B, B\}$, the distribution μ be given by $\mu(1) = 1 - \mu(0) = \frac{1}{2n}$, and the feature map $\phi : \mathcal{X} \times \mathcal{Y}^\star \to [-B, B]$ be given by $\phi(0, \cdot) = 0$, and $\phi(1, y_{1:h}) = y_h$.

Note that under this construction, $\mathbb{P}_{\{x^t\}_{t\in[n]}\sim\pi_0}(x^t = 0 \forall t \in [T]) \geq 1 - n\mu(1) = \frac{1}{2}$. We let E be the event $\{x^t = 0 \forall t \in [T]\}$. Then, for any $\theta^\star \in [-1, 1]$,

$$\mathbb{E}_{\mathcal{D}\sim\pi_{\theta^\star}}[D_{\mathsf{KL}}(\pi_{\theta^\star} \,\|\, \widehat{\pi}) \mid E] = \mathbb{E}_{\mathcal{D}\sim\pi_0}[D_{\mathsf{KL}}(\pi_{\theta^\star} \,\|\, \widehat{\pi}) \mid E].$$

Furthermore, for any $\widehat{\pi} \in \Pi$,

$$D_{\mathsf{KL}}(\pi_{\theta^\star} \,\|\, \widehat{\pi}) = H \cdot \mu(1) \cdot D_{\mathsf{KL}}(\pi_{\theta^\star}(y_1 = \cdot \mid x = 1) \,\|\, \widehat{\pi}(y_1 = \cdot \mid x = 1)),$$

and hence,

$$D_{\mathsf{KL}}(\pi_1 \,\|\, \widehat{\pi}) + D_{\mathsf{KL}}(\pi_{-1} \,\|\, \widehat{\pi}) \geq \frac{H}{2n} \cdot 2D_{\mathsf{KL}}\left(\mathrm{Ber}\left(\frac{e^B}{e^B + e^{-B}}\right) \,\|\, \mathrm{Ber}\left(\frac{1}{2}\right)\right) \geq \frac{H}{2n}.$$

Therefore, we can lower bound

$$\mathbb{E}_{\mathcal{D}\sim\pi_1}[D_{\mathsf{KL}}(\pi_1 \,\|\, \widehat{\pi})] + \mathbb{E}_{\mathcal{D}\sim\pi_{-1}}[D_{\mathsf{KL}}(\pi_{-1} \,\|\, \widehat{\pi})]$$

$$\geq \mathbb{P}(E) \cdot \mathbb{E}_{\mathcal{D}\sim\pi_0}[D_{\mathsf{KL}}(\pi_1 \,\|\, \widehat{\pi}) + D_{\mathsf{KL}}(\pi_{-1} \,\|\, \widehat{\pi}) \mid E] \geq \frac{1}{2} \cdot \frac{H}{2n}.$$

This gives the desired lower bound. □

Note that in the construction above, the variance $\sigma_\star^2$ (defined in Section 4.1) can be bounded by $\sigma_\star^2 \lesssim \frac{He^{-2B}}{n}$.

J Proofs from Section 4

J.1 Proof Sketch for Theorem 4.1

Fix $N \geq 8$ and let $\varepsilon \in [0,1]$ be a parameter to be set later. Let $\Pi_{\mathsf{bad}}(\varepsilon) := \{\pi \in \Pi \mid \mathsf{Cov}_N(\pi) \geq \varepsilon\}$ be the set of $\pi \in \Pi$ that fail to achieve coverage ε. The basic idea behind the proof of Theorem 4.1 is to interpret the condition $\mathsf{Cov}_N(\pi) \geq \varepsilon$ as an small-ball like *anti-concentration* condition in the vein of Mendelson (2014; 2017). That is, for models $\pi \in \Pi_{\mathsf{bad}}(\varepsilon)$ where coverage fails, the condition $\mathsf{Cov}_N(\pi) \geq \varepsilon$ witnesses a *one-sided* tail bound which implies that the empirical likelihood of π is *not too large* with high probability, which means that $\pi \in \Pi_{\mathsf{bad}}(\varepsilon)$ cannot be a maximum-likelihood solution.

Let $\mathcal{S}_N(\pi) := \frac{1}{n}\left|\left\{i \in [n] \mid \frac{\pi_{\mathsf{D}}(y^i|x^i)}{\pi(y^i|x^i)} \geq N^{1-2c}\right\}\right|$ denote the empirical probability of π fails to cover π_{D}. Our first step is to show via covering and concentration that with high-probability, all $\pi \in \Pi$ satisfy

$$\mathcal{S}_N(\pi) \geq \frac{1}{2}\mathsf{Cov}_N(\pi) - \mathcal{C}_{\mathtt{coarse}}(\Pi, N, n), \tag{47}$$

Here we only pays the covering number at a coarse scale (leading to the coarse-grained term in Theorem 4.1) because we only need to show that coverage concentrates, not the log-loss itself.

Eq. (47) implies that for all $\pi \in \Pi_{\mathsf{bad}}(\varepsilon)$, we can bound

$$\begin{aligned}
\widehat{L}_n(\pi) - \widehat{L}_n(\pi_{\mathsf{D}}) &= -\sum_{i=1}^n \left[\log \frac{\pi_{\mathsf{D}}(y^i \mid x^i)}{\pi(y^i \mid x^i)} - C\right]_+ + \sum_{i=1}^n \log \frac{\pi(y^i \mid x^i)}{\pi_{\mathsf{D}}(y^i \mid x^i)} \vee C \\
&\leq -|\mathcal{S}_N(\pi)|((1-2c)\log N - C) + \sum_{i=1}^n \log \frac{\pi(y^i \mid x^i)}{\pi_{\mathsf{D}}(y^i \mid x^i)} \vee C \\
&\leq -\frac{n}{4}\log N \cdot \mathsf{Cov}_N(\pi) + \sum_{i=1}^n \log \frac{\pi(y^i \mid x^i)}{\pi_{\mathsf{D}}(y^i \mid x^i)} \vee C,
\end{aligned}$$

where $C > 0$ is any fixed constant. Finally, using a variation of a standard one-sided tail bound for the logarithmic loss (van de Geer, 2000; Zhang, 2006),[7] we show that with high probability, all $\pi \in \Pi$ satisfy

$$\sum_{i=1}^n \log \frac{\pi(y^i \mid x^i)}{\pi_{\mathsf{D}}(y^i \mid x^i)} \vee C \leq \mathcal{C}_{\mathtt{fine}}(\Pi, n) \cdot n,$$

as long as $C \geq \log 4$. Combining these results, we conclude that

$$\mathsf{Cov}_N(\pi) \lesssim \frac{\widehat{L}_n(\pi_{\mathsf{D}}) - \widehat{L}_n(\pi) + \mathcal{C}_{\mathtt{fine}}(\Pi, n)}{n \log N} + \mathcal{C}_{\mathtt{coarse}}(\Pi, N, n).$$

It follows that if $\varepsilon \gtrsim \frac{1}{\log N} \cdot \mathcal{C}_{\mathtt{fine}}(\Pi, n) + \mathcal{C}_{\mathtt{coarse}}(\Pi, N, n)$, then all $\pi \in \Pi_{\mathsf{bad}}(\varepsilon)$ have $L(\pi) - L(\pi_{\mathsf{D}}) < 0$, and since $\pi_{\mathsf{D}} \in \Pi$, this means that no such $\pi \in \Pi_{\mathsf{bad}}(\varepsilon)$ can be the maximum likelihood solution.

J.2 Proof of Theorem 4.1

Theorem 4.1′ (General version of Theorem 4.1)**.** *Let $N \geq 8$ be given. With probability at least $1-\delta$, any approximate maximum likelihood estimator $\widehat{\pi}$ with $\widehat{L}_n(\widehat{\pi}) \geq \max_{\pi \in \Pi} \widehat{L}_n(\pi) - n\varepsilon_{\mathsf{apx}}$ satisfies*

$$\mathsf{Cov}_N(\widehat{\pi}) \lesssim \frac{\log \mathcal{N}_\infty\left(\Pi; c\log N\right) + \log(\delta^{-1})}{n} + \frac{1}{\log N}\left(\inf_{\varepsilon > 0}\left\{\frac{\log \mathcal{N}_\infty(\Pi, \varepsilon)}{n} + \varepsilon\right\} + \varepsilon_{\mathsf{apx}}\right), \tag{48}$$

where $c > 0$ is an absolute constant.

[7] That the bound is one-sided is critical, as this allows us to avoid paying for the range of the density ratios under consideration. For details, see Proposition J.1.

In the following, for a fixed threshold $C \geq 4$, we define the clipped log loss as

$$L_C^+(\pi) := \sum_{i=1}^n \max\left\{\log \frac{\pi(y^i \mid x^i)}{\pi_{\mathtt{D}}(y^i \mid x^i)}, -\log C\right\}, \tag{49}$$

$$L_C^-(\pi) := \sum_{i=1}^n \max\left\{0, \log \frac{\pi_{\mathtt{D}}(y^i \mid x^i)}{\pi(y^i \mid x^i)} - \log C\right\}. \tag{50}$$

Note that $\widehat{L}_n(\pi) = L_C^+(\pi) - L_C^-(\pi)$, and hence for approximate maximum likelihood estimator $\widehat{\pi}$ with $\widehat{L}_n(\widehat{\pi}) \geq \max_{\pi\in\Pi} \widehat{L}_n(\pi) - n\varepsilon_{\mathsf{apx}}$, we have

$$L_C^-(\widehat{\pi}) \leq L_C^+(\widehat{\pi}) + n\varepsilon_{\mathsf{apx}}.$$

In the following, we argue that $L_C^-(\pi)$ upper bound the coverage $\mathtt{Cov}_N(\pi)$ for any $\pi \in \Pi$ and $M > C$, and $L_C^+(\pi)$ can be bounded by the uniform convergence argument.

Proposition J.1. *Suppose that $C \geq 4$. Then, with probability at least $1-\delta$, it holds that for any $\pi \in \Pi$,*

$$L_C^+(\pi) \leq 2 \inf_{\epsilon \geq 0}\{\log \mathcal{N}_\infty(\Pi, \epsilon) + n\epsilon\}.$$

Proposition J.2. *Fix any $\alpha \in (0, \frac{\log(N/C)}{2})$. Then, with probability at least $1-\delta$, it holds that*

$$\mathtt{Cov}_N(\pi) \leq \frac{2}{\log(N/C) - 2\alpha} \cdot L_C^-(\pi) + 2\log(\mathcal{N}_\infty(\Pi, \alpha)/\delta).$$

The proof of Theorem 4.1 and Theorem 4.1′ is hence completed by combining the propositions above and setting $\alpha = c \log N$. □

Proof of Proposition J.1. This is a direct corollary of Lemma H.4. For each $\pi \in \Pi$, we let $f_\pi(x, y) := \max\left\{\log \frac{\pi(y|x)}{\pi_{\mathtt{D}}(y|x)}, -\log C\right\}$, and then $N(\mathcal{F}, \epsilon; \|\cdot\|_\infty) \leq \mathcal{N}_\infty(\Pi, \epsilon)$ for any $\epsilon \geq 0$. Applying Lemma H.4 with Lemma J.1 gives the desired upper bound. □

Lemma J.1. *As long as $C \geq 4$, it holds that*

$$\mathbb{E}_{(x,y)\sim\pi_{\mathtt{D}}} \exp\left(\frac{1}{2}\max\left\{\log \frac{\pi(y \mid x)}{\pi_{\mathtt{D}}(y \mid x)}, -\log C\right\}\right) \leq 1. \tag{51}$$

Proof of Lemma J.1. We denote $E := \left\{(x, y) : \frac{\pi(y|x)}{\pi_{\mathtt{D}}(y|x)} \geq \frac{1}{C}\right\}$. Then it holds that

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim\pi_{\mathtt{D}}} \exp\left(\frac{1}{2}\max\left\{\log \frac{\pi(y \mid x)}{\pi_{\mathtt{D}}(y \mid x)}, -\log C\right\}\right) \\
&= \mathbb{E}_{(x,y)\sim\pi_{\mathtt{D}}}\left[\sqrt{\frac{\pi(y \mid x)}{\pi_{\mathtt{D}}(y \mid x)}}\mathbb{I}\{(x,y) \in E\} + \frac{1}{\sqrt{C}}\mathbb{I}\{(x,y) \notin E\}\right] \\
&= \mathbb{E}_{x\sim\pi_{\mathtt{D}}}\left[\sum_{y:(x,y)\in E} \sqrt{\pi(y \mid x)\pi_{\mathtt{D}}(y \mid x)}\right] + \frac{1}{\sqrt{C}}\pi_{\mathtt{D}}(E^c)
\end{aligned}$$

By Cauchy inequality, we have

$$\sum_{y:(x,y)\in E} \sqrt{\pi(y \mid x)\pi_{\mathtt{D}}(y \mid x)} \leq \sqrt{\sum_{y:(x,y)\in E} \pi(y \mid x) \cdot \sum_{y:(x,y)\in E} \pi_{\mathtt{D}}(y \mid x)} \leq \sqrt{\pi_{\mathtt{D}}(E)}.$$

Therefore, as long as $C \geq 4$, it holds that

$$\mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}} \exp\left(\frac{1}{2}\max\left\{\log\frac{\pi(y \mid x)}{\pi_{\mathsf{D}}(y \mid x)}, -\log C\right\}\right) \leq \sqrt{\pi_{\mathsf{D}}(E)} + \frac{1}{2}(1-\pi_{\mathsf{D}}(E)) \leq 1,$$

where we use $1-p = (1+\sqrt{p})(1-\sqrt{p}) \leq 2(1-\sqrt{p})$ for any $p \in [0,1]$. □

Proof of Proposition J.2. Fix any $N \geq 1, \alpha \geq 0$. By definition, for any $\pi \in \Pi$,

$$\begin{aligned} L_C^-(\pi) &= \sum_{i=1}^n \max\left\{0, \log\frac{\pi_{\mathsf{D}}(y^i \mid x^i)}{\pi(y^i \mid x^i)} - \log C\right\} \\ &\geq (\log N - \log C)\left|\left\{i \in [n] : \log\frac{\pi_{\mathsf{D}}(y^i \mid x^i)}{\pi(y^i \mid x^i)} \geq \log N\right\}\right| \\ &=: n(\log N - \log C)\cdot \widehat{\mathsf{Cov}}_M(\pi_{\mathsf{D}} \,\|\, \pi), \end{aligned}$$

where we denote (cf. Lemma J.2)

$$\widehat{\mathsf{Cov}}_N(\pi_{\mathsf{D}} \,\|\, \pi) = \frac{1}{n}\left|\left\{t \in [n] : \frac{\pi_{\mathsf{D}}(y^t \mid x^t)}{\pi(y^t \mid x^t)} \geq N\right\}\right|.$$

Then, by Lemma J.2, it holds that with probability at least $1-\delta$, for any $\pi \in \Pi$,

$$\widehat{\mathsf{Cov}}_N(\pi_{\mathsf{D}} \,\|\, \pi) \geq \frac{1}{2}\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{2\alpha}N}(\pi_{\mathsf{D}} \,\|\, \pi) - \log(\mathcal{N}_\infty(\Pi,\alpha)/\delta).$$

Rescaling $N \leftarrow e^{-2\alpha}N$ and reorganizing complete the proof. □

Lemma J.2. *For any policy π, π', we consider the quantities*

$$\widehat{\mathsf{Cov}}_N(\pi' \,\|\, \pi) = \frac{1}{n}\left|\left\{t \in [n] : \frac{\pi'(y^t \mid x^t)}{\pi(y^t \mid x^t)} \geq N\right\}\right|, \qquad \mathsf{Cov}^{\pi_{\mathsf{D}}}_N(\pi' \,\|\, \pi) = \mathbb{P}_{\pi_{\mathsf{D}}}\left(\frac{\pi'(y \mid x)}{\pi(y \mid x)} \geq M\right).$$

Fix $\alpha \geq 0$ and policy $\overline{\pi}$. With probability at least $1-\delta$, for any $\pi \in \Pi$, it holds that

$$\widehat{\mathsf{Cov}}_N(\overline{\pi} \,\|\, \pi) \geq \frac{1}{2}\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{2\alpha}N}(\overline{\pi} \,\|\, \pi) - \log(\mathcal{N}_\infty(\Pi,\alpha)/\delta).$$

Similarly, with probability at least $1-\delta$, for any $\pi \in \Pi$, it holds that

$$\widehat{\mathsf{Cov}}_N(\pi \,\|\, \overline{\pi}) \leq 2\,\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{-2\alpha}N}(\pi \,\|\, \overline{\pi}) + \log(\mathcal{N}_\infty(\Pi,\alpha)/\delta).$$

Proof of Lemma J.2. We only prove the first inequality. Let $\Pi' \subseteq \Pi$ be an α-covering of Π with $|\Pi'| = \mathcal{N}_\infty(\Pi,\alpha)$. Then, by Freedman inequality (Lemma H.3) and union bound, it holds that with probability at least $1-\delta$, for any $\pi' \in \Pi'$,

$$\widehat{\mathsf{Cov}}_{e^\alpha N}(\overline{\pi} \,\|\, \pi') \geq \frac{1}{2}\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^\alpha N}(\overline{\pi} \,\|\, \pi') - \log(|\Pi'|/\delta).$$

Then, note that for any $\pi \in \Pi$, there exists $\pi' \in \Pi'$ such that $|\log\pi(y \mid x) - \log\pi'(y \mid x)| \leq \alpha$ for $\forall x, y$, we know

$$\left\{t \in [n] : \frac{\overline{\pi}(y^t \mid x^t)}{\pi'(y^t \mid x^t)} \geq e^\alpha N\right\} \subseteq \left\{t \in [n] : \frac{\overline{\pi}(y^t \mid x^t)}{\pi(y^t \mid x^t)} \geq N\right\}$$

and hence $\widehat{\mathsf{Cov}}_{e^\alpha N}(\overline{\pi} \,\|\, \pi') \leq \widehat{\mathsf{Cov}}_N(\overline{\pi} \,\|\, \pi)$. Similarly, $\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^\alpha N}(\overline{\pi} \,\|\, \pi') \geq \mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{2\alpha}N}(\overline{\pi} \,\|\, \pi)$. Hence, under the above event, it holds that

$$\begin{aligned} \widehat{\mathsf{Cov}}_N(\overline{\pi} \,\|\, \pi) \geq \widehat{\mathsf{Cov}}_{e^\alpha N}(\overline{\pi} \,\|\, \pi') &\geq \frac{1}{2}\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^\alpha N}(\overline{\pi} \,\|\, \pi') - \log(|\Pi'|/\delta) \\ &\geq \frac{1}{2}\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{2\alpha}N}(\overline{\pi} \,\|\, \pi) - \log(|\Pi'|/\delta). \end{aligned}$$

Since $\pi \in \Pi$ is arbitrary, the proof is hence completed. □

J.3 Proof of Theorem F.1

By definition and concavity of $\theta \mapsto \pi(y \mid x)$, we know $\theta^\star$ is the optimal solution of the following convex problem

$$\theta^\star = \arg\min_{\theta \in \Theta} - \mathbb{E}_{(x,y)\sim\pi_{\mathrm{D}}}[\log \pi_\theta(y \mid x)].$$

Hence, the optimality of $\theta^\star$ implies

$$\langle \theta - \theta^\star, - \mathbb{E}_{\pi_{\mathrm{D}}}[\nabla \log \pi_{\theta^\star}(y \mid x)] \rangle \geq 0, \qquad \forall \theta \in \Theta.$$

Consider the function $F(\theta) = \mathbb{E}_{\pi_{\mathrm{D}}}\left[\frac{\pi_\theta(y|x)}{\pi_{\theta^\star}(y|x)}\right] - 1$, which is also convex by Assumption F.1. Further, for any $\theta \in \Theta$,

$$\langle \theta - \theta^\star, -\nabla F(\theta^\star) \rangle = \left\langle \theta - \widehat{\theta}, - \mathbb{E}_{\pi_{\mathrm{D}}}\left[\frac{\nabla \pi_{\theta^\star}(y \mid x)}{\pi_{\theta^\star}(y \mid x)}\right] \right\rangle = \left\langle \theta - \widehat{\theta}, - \mathbb{E}_{\pi_{\mathrm{D}}}[\nabla \log \pi_{\theta^\star}(y \mid x)] \right\rangle \geq 0.$$

Therefore, F attains its maximum over Θ at $\theta^\star$, i.e., $F(\theta) \leq F(\theta^\star)$ for any $\theta \in \Theta$.

Similarly, under Assumption F.1, it is clear that $\theta \mapsto \sum_{i=1}^n \log \pi_\theta(y^i \mid x^i)$ is concave, and hence $\widehat{\pi} = \pi_{\widehat{\theta}}$, where $\widehat{\theta} \in \Theta$ satisfies

$$\left\langle \theta - \widehat{\theta}, \sum_{i=1}^n -\nabla \log \pi_{\widehat{\theta}}(y^i \mid x^i) \right\rangle \geq 0, \qquad \forall \theta \in \Theta.$$

In particular, we consider the function

$$\widehat{F}(\theta) := \sum_{i=1}^n \left[\frac{\pi_\theta(y^i \mid x^i)}{\pi_{\widehat{\theta}}(y^i \mid x^i)} - 1\right].$$

By definition, $\widehat{F}$ is concave, and for any $\theta \in \Theta$,

$$\left\langle \theta - \widehat{\theta}, -\nabla \widehat{F}(\widehat{\theta}) \right\rangle = \left\langle \theta - \widehat{\theta}, - \sum_{i=1}^n \frac{\nabla \pi_{\widehat{\theta}}(y^i \mid x^i)}{\pi_{\widehat{\theta}}(y^i \mid x^i)} \right\rangle = \left\langle \theta - \widehat{\theta}, \sum_{i=1}^n -\nabla \log \pi_{\widehat{\theta}}(y^i \mid x^i) \right\rangle \geq 0.$$

Therefore, $\widehat{F}$ attains its maximum over Θ at $\widehat{\theta}$, and in particular, $\widehat{F}(\theta^\star) \leq \widehat{F}(\widehat{\theta}) = 0$. This implies

$$\sum_{i=1}^n \left[\frac{\pi_{\theta^\star}(y^i \mid x^i)}{\widehat{\pi}(y^i \mid x^i)} - \log \frac{\pi_{\theta^\star}(y^i \mid x^i)}{\widehat{\pi}(y^i \mid x^i)} - 1\right] \leq \sum_{i=1}^n \log \widehat{\pi}(y^i \mid x^i) - \sum_{i=1}^n \log \pi_{\theta^\star}(y^i \mid x^i). \tag{52}$$

In the following, we fix any $N \geq 2$. Note that $x - \log x - 1 \geq 0$ for any $x > 0$, and $x \mapsto x - \log x - 1$ is increasing for $x \geq 1$. Therefore, (52) implies that

$$(N - \log N - 1) \cdot n \cdot \widehat{\mathsf{Cov}}_N(\pi_{\theta^\star} \,\|\, \widehat{\pi}) \leq \widehat{L}_n(\widehat{\pi}) - \widehat{L}_n(\pi_{\theta^\star}). \tag{53}$$

Then, by Lemma J.2, we have with probability at least $1 - \delta$, for all $\pi \in \Pi$,

$$\widehat{\mathsf{Cov}}_N(\pi_{\theta^\star} \,\|\, \pi) \geq \frac{1}{2} \cdot \mathbb{P}_{\pi_{\mathrm{D}}}\left(\frac{\pi_{\theta^\star}(y \mid x)}{\pi(y \mid x)} \geq e^{2\alpha} N\right) - \frac{\log(\mathcal{N}_\infty(\Pi, \alpha)/\delta)}{n}, \qquad \forall \pi \in \Pi.$$

Further, by Lemma H.4, the following holds with probability at least $1 - \delta$: For any $\theta \in \Theta$,

$$\begin{aligned}
\widehat{L}_n(\pi_\theta) - \widehat{L}_n(\pi_{\theta^\star}) &= \sum_{i=1}^n \log \frac{\pi_\theta(y^i \mid x^i)}{\pi_{\theta^\star}(y^i \mid x^i)} \\
&\leq n \log \mathbb{E}_{\pi_{\mathrm{D}}}\left[\frac{\pi_\theta(y \mid x)}{\pi_{\theta^\star}(y \mid x)}\right] + \inf_{\epsilon \geq 0}\{\log(\mathcal{N}_\infty(\Pi, \epsilon)/\delta) + 2n\epsilon\} \\
&\leq \inf_{\epsilon \geq 0}\{\log(\mathcal{N}_\infty(\Pi, \epsilon)/\delta) + 2n\epsilon\},
\end{aligned}$$

where we use $\mathbb{E}_{\pi_{\mathrm{D}}}\left[\frac{\pi_\theta(y|x)}{\pi_{\theta^\star}(y|x)}\right] = F(\theta) + 1 \leq 1$ for any $\theta \in \Theta$. By union bound, we have shown that with probability at least $1 - 2\delta$,

$$\mathbb{P}_{\pi_{\mathrm{D}}}\left(\frac{\pi_{\theta^\star}(y \mid x)}{\widehat{\pi}(y \mid x)} \geq e^{2\alpha} N\right) \lesssim \frac{\log(\mathcal{N}_\infty(\Pi, \alpha)/\delta)}{n} + \frac{1}{N} \inf_{\epsilon \geq 0}\left\{\frac{\log \mathcal{N}_\infty(\Pi, \epsilon)}{n} + \epsilon\right\}.$$

Note that

$$\begin{aligned}\mathsf{Cov}_{e^{2\alpha}NN'}(\widehat{\pi}) &= \mathbb{P}_{\pi_{\mathsf{D}}}\left(\frac{\pi_{\mathsf{D}}(y \mid x)}{\widehat{\pi}(y \mid x)} \geq e^{2\alpha}NN'\right) \\ &\leq \mathbb{P}_{\pi_{\mathsf{D}}}\left(\frac{\pi_{\theta^\star}(y \mid x)}{\pi(y \mid x)} \geq e^{2\alpha}N\right) + \mathbb{P}_{\pi_{\mathsf{D}}}\left(\frac{\pi_{\mathsf{D}}(y \mid x)}{\pi_{\theta^\star}(y \mid x)} \geq N'\right).\end{aligned}$$

Therefore, the proof is completed by rescaling $N \leftarrow Ne^{-2\alpha}/N'$ and combining the inequalities above.

J.4 Proofs for Supporting Results

Proof of Proposition F.1 (a). Assume that $B \geq \frac{1}{2}\log(4n)$ and $n \geq d$. Consider $\mathcal{X} = \perp$, $\mathcal{Y} = [d]$ and the feature map be given by $\phi(y) = Be_y$ for $y \in \mathcal{Y}$, where $(e_1, \cdots, e_d)$ is the coordinate basis of $\mathbb{R}^d$. For the simplicity of our argument, we consider $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_\infty \leq 1\right\}$, and we set

$$\theta^\star = \frac{\log(4n)}{2B} \cdot \left(e_1 - \sum_{j=2}^{d} e_j\right)$$

Then it holds that

$$\pi_{\mathsf{D}}(1) = \frac{4n}{d-1+4n}, \qquad \pi_{\mathsf{D}}(y) = \frac{1}{d-1+4n}, \qquad \forall y > 1.$$

Therefore, under $\mathcal{D} \sim \pi_{\mathsf{D}}$, it holds that

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\{y^t \neq 1\}\right] \leq \frac{n(d-1)}{d-1+4n} \leq \frac{d-1}{4} \leq \frac{n}{2}.$$

In particular, with probability at least 0.5, it holds that $\sum_{t=1}^{T} \mathbb{I}\{y^t \neq 1\} \leq \frac{d-1}{2}$. i.e., the set $\mathcal{Y}_{\mathcal{D}} = \mathcal{Y} \backslash \mathcal{D}$ has cardinality at least $\frac{d}{2}$.

In the following, we condition on this event and analyze the MLE $\widehat{\theta}$. By the definition of MLE, for any $y \in \mathcal{Y}_{\mathcal{D}}$, it must hold that $\widehat{\theta}_y = -1$, and we also know $\widehat{\theta}_1 = 1$. This implies $\pi_{\widehat{\theta}}(y) \leq \frac{1}{e^{2B}}$ for any $y \in \mathcal{Y}_{\mathcal{D}}$. Therefore, for $N \leq \frac{e^B}{4n+d-1}$, we have

$$\mathsf{Cov}_N(\pi_{\widehat{\theta}}) \geq \pi_{\mathsf{D}}(\mathcal{Y}_{\mathcal{D}}) \geq \frac{d}{2(d-1+4n)} \geq \frac{d}{10n}.$$

This is the desired lower bound. □

Proof of Proposition F.1 (b). Let $\epsilon = c_0\sqrt{\frac{d}{n}}$ and $p = \frac{\epsilon^2}{\log N}$ for a sufficently small absolute constant $c_0 > 0$. Let $\mathcal{X} = \{0, 1, \cdots, d\}$, $\mathcal{Y} = \{0, 1\}$, and the distribution μ be given by $\mu(0) = p$, $\mu(1) = \cdots = \mu(d) = \frac{1-p}{d}$.

Consider $\pi_{\mathsf{D}}(\cdot \mid i) = \mathrm{Ber}(1/2)$ for $i \in [d]$ and $\pi_{\mathsf{D}}(1|0) = 1$. For any $\theta \in \Theta := \{+1, -1\}^d$, we define π_θ as

$$\pi_\theta(1|0) = \frac{1}{N}, \qquad \pi_\theta(\cdot|i) = \mathrm{Ber}\left(\frac{1+\epsilon\theta_i}{2}\right), \qquad \forall i \in [d].$$

Then, we can calculate

$$\begin{aligned}&\max_{\theta\in\Theta} \widehat{L}_n(\pi_\theta) - \widehat{L}_n(\pi_{\mathsf{D}}) \\ =\ & -N(0)\log N + \frac{1}{2}\sum_{i\in[d]}\left[|N(i,+) - N(i,-)|\log\frac{1+\epsilon}{1-\epsilon} + N(i)\log(1-\epsilon^2)\right] \\ \geq\ & -N(0)\log N - n\epsilon^2 + \frac{\epsilon}{2}\sum_{i\in[d]}|N(i,0) - N(i,1)|,\end{aligned}$$

where we denote $N(x,y) = \#\{t \in [n] : (x^t, y^t) = (x,y)\}$, $N(x) = N(x,0) + N(x,1)$. In the following, we denote $\Delta_i = N(i,1) - N(i,0)$. Note that Δ_i is a sum of n i.i.d $\{-1,0,1\}$-valued random variables with mean zero and variance $\frac{1-p}{d}$, and hence $\mathbb{P}_{\pi_{\mathsf{D}}}\big(|\Delta_i| \geq c\sqrt{\frac{n}{d}}\big) \geq 0.99$ for an absolute constant $c > 0$. Then, it is clear that

$$\mathbb{P}_{\pi_{\mathsf{D}}}\left(\sum_{i=1}^{d}|\Delta_i| \leq \frac{d}{2}\cdot c\sqrt{\frac{n}{d}}\right) \leq \frac{1}{4}.$$

By Markov inequality, we also know $\mathbb{P}_{\pi_{\mathsf{D}}}(N(0) \geq 4np) \leq \frac{1}{4}$. Combining the inequalities above, we know with probability at least 0.5, we have

$$\max_{\theta\in\Theta} L(\pi_\theta) - L(\pi_{\mathsf{D}}) \geq -4np\log N - n\epsilon^2 + \frac{c\epsilon\sqrt{nd}}{4} > 0.$$

This implies that the MLE $\widehat{\pi} \in \{\pi_\theta\}_{\theta\in\Theta}$, and hence $\mathtt{Cov}_N(\widehat{\pi}) \geq p$. This is the desired lower bound. □

K Proofs from Section 5

Organization. We begin with the proof of Proposition 5.1 (the upper bound), which is relatively simple and serves as motivation. We then present the proofs of Theorem 4.2 and Theorem 5.1, which are more involved. Finally, the proofs of the lower bounds are given in the remaining subsections.

Notation. For notational simplicity, we denote

$$\bar{\phi}_\theta(x, y_{1:h-1}) = \mathbb{E}_{y_h\sim\pi_\theta(\cdot|x,y_{1:h-1})}[\phi(x, y_{1:h})],$$

and

$$\phi^\star(x, y_{1:h}) := \phi(x, y_{1:h}) - \bar{\phi}_{\theta^\star}(x, y_{1:h-1}),$$
$$\mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1}) := \mathbb{E}_{y_h\sim\pi_\theta(\cdot|x,y_{1:h-1})}\|\phi^\star(x, y_{1:h})\|^2.$$

Then, by definition,

$$\nabla\log\pi_\theta(y_{1:H} \mid x) = \sum_{h=1}^{H}\phi^\star(x, y_{1:h}) + \sum_{h=1}^{H}\big(\bar{\phi}_{\theta^\star}(x, y_{1:h-1}) - \bar{\phi}_\theta(x, y_{1:h-1})\big). \tag{54}$$

We also write

$$\epsilon_\theta(x, y_{1:h-1}) = D_{\mathsf{KL}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}) \,\|\, \pi_\theta(\cdot \mid x, y_{1:h-1})).$$

By concavity, we have

$$\epsilon_\theta(x, y_{1:h-1}) \leq \langle\bar{\phi}_\theta(x, y_{1:h-1}) - \bar{\phi}_{\theta^\star}(x, y_{1:h-1}), \theta - \theta^\star\rangle. \tag{55}$$

By Lemma H.5, it holds that

$$\|\bar{\phi}_{\theta^\star}(x, y_{1:h-1}) - \bar{\phi}_\theta(x, y_{1:h-1})\| \leq 2\sqrt{\mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\cdot\epsilon_\theta(x, y_{1:h-1})} + 3B\epsilon_\theta(x, y_{1:h-1}). \tag{56}$$

For notational simplicity, for any $f : \mathcal{X}\times\mathcal{A}^\star \to \mathbb{R}$ and dataset $\mathcal{D} = \{(x^i, y^i_{1:H})\}_{i\in[n]}$, we write

$$\widehat{\mathbb{E}}_{\mathcal{D}}[f] := \frac{1}{n}\sum_{i=1}^{n} f(x^i, y^i_{1:H}),$$

K.1 Proof of Proposition 5.1 (Upper Bound)

Because the projection operator Proj_Θ is an contraction, we have

$$\begin{aligned}
&\|\theta^t - \theta^\star\|^2 - \|\theta^{t+1} - \theta^\star\|^2 \\
&\geq \|\theta^t - \theta^\star\|^2 - \|\theta^t + \eta\nabla\log\pi_{\theta^t}(y^t \mid x^t) - \theta^\star\|^2 \\
&= 2\eta\langle-\nabla\log\pi_{\theta^t}(y^t \mid x^t), \theta^t - \theta^\star\rangle - 2\eta^2\|\nabla\log\pi_{\theta^t}(y^t \mid x^t)\|^2.
\end{aligned}$$

Telescoping and taking expectation, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle -\nabla \log \pi_{\theta^t}(y \mid x), \theta^t - \theta^\star\rangle\right] \le \frac{1}{2\eta} + \eta\,\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}\|\nabla \log \pi_{\theta^t}(y \mid x)\|^2\right]. \tag{57}$$

Note that $(x^t, y^t) \mid \theta^t \sim \pi_{\mathsf{D}}$, and hence

$$\mathbb{E}[\nabla \log \pi_{\theta^t}(y^t \mid x^t) \mid \theta^t] = \mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}[\nabla \log \pi_{\theta^t}(y \mid x)] = \nabla_\theta D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \pi_\theta)|_{\theta=\theta^t}.$$

Further, by convexity, it holds that for any $\theta \in \Theta$,

$$G(\theta) := \mathbb{E}_{\pi_{\mathsf{D}}}[\langle \nabla \log \pi_\theta(y \mid x), \theta - \theta^\star\rangle] = \langle \nabla_\theta D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \pi_\theta), \theta - \theta^\star\rangle \ge D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \pi_\theta).$$

Therefore, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \pi_{\theta^t})\right] \le \mathbb{E}\left[\sum_{t=1}^{T} G(\theta^t)\right] \le \frac{1}{2\eta} + \eta\,\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}\|\nabla \log \pi_{\theta^t}(y \mid x)\|^2\right].$$

On the other hand, using the fact that $\log \pi_\theta(y \mid x)$ is concave and (HB^2)-smooth (i.e., $-HB^2 I \preceq \nabla^2 \log \pi_\theta(y \mid x) \preceq 0$),

$$\|\nabla \log \pi_\theta(y \mid x) - \nabla \log \pi_{\theta^\star}(y \mid x)\|^2 \le HB^2 \cdot \langle \theta - \theta^\star, \nabla \log \pi_{\theta^\star}(y \mid x) - \nabla \log \pi_\theta(y \mid x)\rangle$$

Taking expectation of $(x, y) \sim \pi_{\mathsf{D}}$ and using the fact that $\mathbb{E}_{\pi_{\mathsf{D}}}[\nabla \log \pi_{\theta^\star}(y \mid x)] = 0$, we have

$$\mathbb{E}_{\pi_{\mathsf{D}}}\|\nabla \log \pi_\theta(y \mid x) - \nabla \log \pi_{\theta^\star}(y \mid x)\|^2 \le HB^2 \cdot G(\theta), \qquad \forall \theta \in \Theta.$$

Further, note that $\mathbb{E}_{\pi_{\mathsf{D}}}\|\nabla \log \pi_{\theta^\star}(y \mid x)\|^2 = \sigma_\star^2$, it holds that

$$\mathbb{E}_{\pi_{\mathsf{D}}}\|\nabla \log \pi_{\theta^\star}(y \mid x)\|^2 \le 2\sigma_\star^2 + 2HB^2 \cdot G(\theta), \qquad \forall \theta \in \Theta. \tag{58}$$

Combining the inequalities above, we can conclude that

$$\mathbb{E}\left[\sum_{t=1}^{T} G(\theta^t)\right] \le \frac{1}{2\eta} + 2\eta HB^2\,\mathbb{E}\left[\sum_{t=1}^{T} G(\theta^t)\right] + 2\eta T\sigma_\star^2.$$

Therefore, as long as $\eta \le \frac{1}{4HB^2}$, it holds

$$\frac{1}{\eta} + 4\eta T\sigma_\star^2 \ge \mathbb{E}\left[\sum_{t=1}^{T} G(\theta^t)\right] \ge \mathbb{E}\left[\sum_{t=1}^{T} D_{\mathsf{KL}}(\pi_{\mathsf{D}} \,\|\, \pi_{\theta^t})\right].$$

This is the desired upper bound. □

K.2 Proof of Theorem 5.1

We denote $M := \log N$, and we analyze the normalized SGD iterates assuming $\lambda \ge 3BM$ and $\frac{\lambda\eta}{M} \le \frac{1}{8}$. and

Denote

$$\widetilde{g}_\theta(\mathcal{D}) := \frac{\widehat{g}_\theta(\mathcal{D})}{\lambda + \|\widehat{g}_\theta(\mathcal{D})\|}.$$

Then the normalized SGD update can be rewritten as $\theta^{t+1} = \mathrm{Proj}_\Theta(\theta + \eta\widetilde{g}_{\theta^t}(\mathcal{D}))$. By the standard SGD proof, we know that

$$\sum_{t=1}^{T}\langle -\widetilde{g}_{\theta^t}(\mathcal{D}), \theta^t - \theta^\star\rangle \le \frac{\|\theta^0 - \theta^\star\|^2}{2\eta} + \eta\sum_{t=1}^{T}\|\widetilde{g}_{\theta^t}(\mathcal{D})\|^2.$$

Taking expectation on both sides, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}\sim\pi_{\mathsf{D}}}\langle -\widetilde{g}(\theta^t; \mathcal{D}), \theta^t - \theta^\star\rangle\right] \le \frac{\|\theta^0 - \theta^\star\|^2}{2\eta} + \eta\,\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\mathcal{D}\sim\pi_{\mathsf{D}}}\|\widetilde{g}(\theta^t; \mathcal{D})\|^2\right].$$

Note that $\|\widetilde{g}_\theta(\mathcal{D})\| \le \min\Big\{1, \frac{\|\widehat{g}_\theta(\mathcal{D})\|}{\lambda}\Big\}$. In the following, we analyze $\|\widehat{g}_\theta(\mathcal{D})\|$ under $\mathcal{D} = \{(x^i, y^i_{1:H})\}_{i\in[K]} \sim \pi_{\mathsf{D}}$. Recall that for notational simplicity, for any $f : \mathcal{X} \times \mathcal{A}^\star \to \mathbb{R}$, we write

$$\widehat{\mathbb{E}}_{\mathcal{D}}[f] := \frac{1}{K}\sum_{i=1}^{K} f(x^i, y^i_{1:H}),$$

which is random variable. We denote

$$\bar{g}_\theta(\mathcal{D}) := \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\big(\bar{\phi}_\theta(x, y_{1:h-1}) - \bar{\phi}_{\theta^\star}(x, y_{1:h-1})\big)\right],$$

and

$$z(\mathcal{D}) := \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\phi^\star(x, y_{1:h})\right] = \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\big(\phi(x, y_{1:h}) - \bar{\phi}_{\theta^\star}(x, y_{1:h-1})\big)\right].$$

Then, by definition, $-\widehat{g}_\theta(\mathcal{D}) = \bar{g}_\theta(\mathcal{D}) - z(\mathcal{D})$.

Bounds on $\bar{g}_\theta(\mathcal{D})$. By (55), we know

$$\begin{aligned}\langle \bar{g}_\theta(\mathcal{D}), \theta - \theta^\star\rangle &= \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\big\langle \bar{\phi}_\theta(x, y_{1:h-1}) - \bar{\phi}_{\theta^\star}(x, y_{1:h-1}), \theta - \theta^\star\big\rangle\right] \\ &\ge \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\epsilon_\theta(x, y_{1:h-1})\right] =: \epsilon_\theta(\mathcal{D}).\end{aligned}$$

By (56), we also have

$$\begin{aligned}\|\bar{g}_\theta(\mathcal{D})\| &\le \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\|\bar{\phi}_\theta(x, y_{1:h-1}) - \bar{\phi}_{\theta^\star}(x, y_{1:h-1})\|\right] \\ &\le \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H} 2\sqrt{\mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\cdot\epsilon_\theta(x, y_{1:h-1})} + 3B\epsilon_\theta(x, y_{1:h-1})\right] \\ &\le 2\sqrt{\sigma^2(\mathcal{D})\cdot\epsilon_\theta(\mathcal{D})} + 3B\epsilon_\theta(\mathcal{D}),\end{aligned}$$

where we denote

$$\sigma^2(\mathcal{D}) := \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\right].$$

Bounds on $z(\mathcal{D})$. Note that $K\cdot z(\mathcal{D}) = K\cdot\widehat{\mathbb{E}}_{\mathcal{D}}\Big[\sum_{h=1}^{H}\phi^\star(x, y_{1:h})\Big] = \sum_{i=1}^{K}\sum_{h=1}^{H}\phi^\star(x^i, y^i_{1:h})$ is a sum of the martingale difference sequence $\{\phi^\star(x^i, y^i_{1:h})\}_{i\in[K], h\in[H]}$. Therefore, we can calculate

$$\mathbb{E}_{\pi_{\mathsf{D}}}\|z(\mathcal{D})\|^2 = \frac{1}{K}\,\mathbb{E}_{\pi_{\mathsf{D}}}\left[\sum_{h=1}^{H}\|\phi^\star(x, y_{1:h})\|^2\right] = \frac{\sigma_\star^2}{K}.$$

Furthermore, by Freedman's inequality (Lemma H.2), for any fixed vector v, parameter $\gamma \in (0, \frac{1}{B})$ and $\delta \in (0,1)$, it holds that

$$\mathbb{P}_{\pi_{\mathsf{D}}}\left(\sum_{i=1}^{K}\sum_{h=1}^{H}\big(\langle\phi^\star(x^i, y^i_{1:h}), v\rangle - \gamma\,\mathbb{E}\big[\langle\phi^\star(x^i, y^i_{1:h}), v\rangle^2 \mid x^i, y^i_{1:h-1}\big]\big) \ge \gamma^{-1}\log(1/\delta)\right) \le \delta.$$

Note that for $v = \theta - \theta^\star$, by Lemma H.6, we have

$$\begin{aligned}&\mathbb{E}\big[\langle\phi^\star(x^i, y^i_{1:h}), v\rangle^2 \mid x^i, y^i_{1:h-1}\big] \\ &= \mathbb{E}_{y_h\sim\pi_{\mathsf{D}}(\cdot\mid x^i, y^i_{1:h-1})}\langle\phi^\star(x^i, y^i_{1:h-1}, y_h), v\rangle^2 \\ &\le 15BD_{\mathsf{KL}}\big(\pi_{\mathsf{D}}(\cdot\mid x^i, y^i_{1:h-1})\,\|\,\pi_\theta(\cdot\mid x^i, y^i_{1:h-1})\big) = 15B\epsilon_\theta(x^i, y^i_{1:h-1}).\end{aligned}$$

Therefore, setting $\gamma = \frac{1}{30B}$, we have shown that for any $\delta \in (0,1)$, it holds that

$$\mathbb{P}_{\pi_{\mathrm{D}}}\left(\langle z(\mathcal{D}), \theta - \theta^\star\rangle \geq \frac{1}{2}\widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\epsilon_\theta(x, y_{1:h-1})\right] + \frac{30B\log(1/\delta)}{K}\right) \leq \delta.$$

Recall that we denote $\epsilon_\theta(\mathcal{D}) := \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\epsilon_\theta(x, y_{1:h-1})\right]$. Therefore, taking integration gives

$$\mathbb{E}_{\pi_{\mathrm{D}}}\left(\langle z(\mathcal{D}), \theta - \theta^\star\rangle - \frac{1}{2}\epsilon_\theta(\mathcal{D})\right)_+ \leq \frac{30B}{K} =: \alpha_K.$$

Upper bounding $\|\widetilde{g}_\theta(\mathcal{D})\|$. Using $\|\widetilde{g}_\theta(\mathcal{D})\| \leq \min\left\{1, \frac{\|\widehat{g}_\theta(\mathcal{D})\|}{\lambda}\right\}$, we know

$$\|\widetilde{g}_\theta(\mathcal{D})\|^2 \leq \mathbb{I}\{\epsilon_\theta(\mathcal{D}) \geq M\} + \mathbb{I}\{\epsilon_\theta(\mathcal{D}) \leq M\} \cdot \frac{\|\widehat{g}_\theta(\mathcal{D})\|}{\lambda}.$$

Therefore, for any fixed θ, it holds that

$$\begin{aligned}
&\mathbb{E}_{\pi_{\mathrm{D}}}\|\widetilde{g}_\theta(\mathcal{D})\|^2 \\
&\leq \mathbb{P}_{\pi_{\mathrm{D}}}(\epsilon_\theta(\mathcal{D}) \geq M) + \frac{1}{\lambda}\,\mathbb{E}_{\pi_{\mathrm{D}}}[\mathbb{I}\{\epsilon_\theta(\mathcal{D}) \leq M\} \cdot \|\bar{g}_\theta(\mathcal{D})\|] + \frac{1}{\lambda}\,\mathbb{E}_{\pi_{\mathrm{D}}}\|z(\mathcal{D})\| \\
&\leq \max\left\{\frac{1}{M}, \frac{3B}{\lambda}\right\} \cdot \mathbb{E}_{\pi_{\mathrm{D}}}\min\{M, \epsilon_\theta(\mathcal{D})\} + \frac{2}{\lambda}\,\mathbb{E}_{\pi_{\mathrm{D}}}\sqrt{\sigma^2(\mathcal{D}) \cdot \min\{M, \epsilon_\theta(\mathcal{D})\}} + \frac{\sigma_\star}{\lambda\sqrt{K}} \\
&\leq \frac{1}{M}\,\mathbb{E}_{\pi_{\mathrm{D}}}\min\{M, \epsilon_\theta(\mathcal{D})\} + \frac{\sigma_\star}{\lambda}\left[2\sqrt{\mathbb{E}_{\pi_{\mathrm{D}}}\min\{M, \epsilon_\theta(\mathcal{D})\}} + \frac{1}{\sqrt{K}}\right],
\end{aligned}$$

where the last inequality follows from $\lambda \geq 3BM$, $\mathbb{E}[\sigma^2(\mathcal{D})] = \sigma_\star^2$, and Cauchy's inequality.

Lower bounding $\langle -\widetilde{g}_\theta(\mathcal{D}), \theta - \theta^\star\rangle$. By the inequalities above, we know

$$\begin{aligned}
\Lambda_\theta &:= \mathbb{E}_{\pi_{\mathrm{D}}}\langle -\widetilde{g}_\theta(\mathcal{D}), \theta - \theta^\star\rangle \\
&= \mathbb{E}_{\pi_{\mathrm{D}}}\left[\frac{\langle \bar{g}_\theta(\mathcal{D}), \theta - \theta^\star\rangle - \langle z(\mathcal{D}), \theta - \theta^\star\rangle}{\lambda + \|\widehat{g}_\theta(\mathcal{D})\|}\right] \\
&\geq \mathbb{E}_{\pi_{\mathrm{D}}}\left[\frac{\epsilon_\theta(\mathcal{D}) - \langle z(\mathcal{D}), \theta - \theta^\star\rangle}{\lambda + \|\widehat{g}_\theta(\mathcal{D})\|}\right] \\
&\geq \frac{1}{2}\,\mathbb{E}_{\pi_{\mathrm{D}}}\left[\frac{\epsilon_\theta(\mathcal{D})}{\lambda + \|\widehat{g}_\theta(\mathcal{D})\|}\right] - \frac{1}{\lambda}\,\mathbb{E}_{\pi_{\mathrm{D}}}\left[\left(\langle z(\mathcal{D}), \theta - \theta^\star\rangle - \frac{1}{2}\epsilon_\theta(\mathcal{D})\right)_+\right] \\
&\geq \frac{1}{2}\,\mathbb{E}_{\pi_{\mathrm{D}}}\left[\frac{\epsilon_\theta(\mathcal{D})}{\lambda + \|z(\mathcal{D})\| + \|\bar{g}_\theta(\mathcal{D})\|}\right] - \frac{\alpha_K}{\lambda}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\lambda + \|z(\mathcal{D})\| + \|\bar{g}_\theta(\mathcal{D})\| \\
&\leq \lambda + \|z(\mathcal{D})\| + 2\sqrt{\sigma^2(\mathcal{D}) \cdot \epsilon_\theta(\mathcal{D})} + 3B\epsilon_\theta(\mathcal{D}) \\
&\leq \frac{\max\{M, \epsilon_\theta(\mathcal{D})\}}{M} \cdot \left[2\lambda + \|z(\mathcal{D})\| + 2M\sqrt{\frac{\sigma^2(\mathcal{D})}{\min\{M, \epsilon_\theta(\mathcal{D})\}}}\right],
\end{aligned}$$

where we use $\min\{M, x\}\max\{M, x\} = Mx$, and $\lambda \geq 3BM$. Combining these two inequalities, we have

$$\begin{aligned}
2\Lambda_\theta + \frac{2\alpha_K}{\lambda} &\geq \mathbb{E}_{\pi_{\mathrm{D}}}\left[\frac{\epsilon_\theta(\mathcal{D})}{\lambda + \|z(\mathcal{D})\| + \|\bar{g}_\theta(\mathcal{D})\|}\right] \\
&\geq \mathbb{E}_{\pi_{\mathrm{D}}}\left[\frac{\min\{M, \epsilon_\theta(\mathcal{D})\}}{2\lambda + \|z(\mathcal{D})\| + 2M\sqrt{\sigma^2(\mathcal{D})/\min\{M, \epsilon_\theta(\mathcal{D})\}}}\right] \\
&\geq \frac{\left(\mathbb{E}_{\pi_{\mathrm{D}}}\min\{M, \epsilon_\theta(\mathcal{D})\}\right)^2}{\mathbb{E}_{\pi_{\mathrm{D}}}[\min\{M, \epsilon_\theta(\mathcal{D})\}(2\lambda + \|z(\mathcal{D})\|)] + 2M\sqrt{\mathbb{E}_{\pi_{\mathrm{D}}}\sigma^2(\mathcal{D}) \cdot \mathbb{E}_{\pi_{\mathrm{D}}}\min\{M, \epsilon_\theta(\mathcal{D})\}}} \\
&\geq \frac{\left(\mathbb{E}_{\pi_{\mathrm{D}}}\min\{M, \epsilon_\theta(\mathcal{D})\}\right)^2}{2\lambda\,\mathbb{E}_{\pi_{\mathrm{D}}}\min\{M, \epsilon_\theta(\mathcal{D})\} + M\sqrt{\sigma_\star^2/K} + 2M\sqrt{\sigma_\star^2\,\mathbb{E}_{\pi_{\mathrm{D}}}\min\{M, \epsilon_\theta(\mathcal{D})\}}},
\end{aligned}$$

where the last two inequalities follow from Cauchy's inequality.

Putting everything together. Denote $\Delta_\theta := \mathbb{E}_{\pi_{\mathtt{D}}} \min\{M, \epsilon_\theta(\mathcal{D})\}$. Note that by Proposition D.10, we have (recall $M := \log N$)

$$\begin{aligned}\mathtt{Cov}_N(\pi_\theta) &\le \frac{1}{M}\mathbb{E}_{\pi_{\mathtt{D}}} \min\left\{M, \sum_{h=1}^H \epsilon_\theta(x, y_{1:h-1})\right\} \\ &\le \frac{1}{M}\mathbb{E}_{\mathcal{D}\sim\pi_{\mathtt{D}}} \min\left\{M, \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^H \epsilon_\theta(x, y_{1:h-1})\right]\right\} = \frac{1}{M}\Delta_\theta.\end{aligned}$$

Then, we have shown that for any fixed parameter θ,

$$\mathbb{E}_{\pi_{\mathtt{D}}}\|\widetilde{g}_\theta(\mathcal{D})\|^2 \le \frac{1}{M}\Delta_\theta + \frac{\sigma_\star}{\lambda}\left[2\sqrt{\Delta_\theta} + \frac{1}{\sqrt{K}}\right],$$

and

$$\Lambda_\theta = \mathbb{E}_{\pi_{\mathtt{D}}}\langle -\widetilde{g}_\theta(\mathcal{D}), \theta - \theta^\star\rangle \ge \frac{1}{2}\frac{\Delta_\theta^2}{2\lambda\Delta_\theta + M\sigma_\star\left[\frac{1}{\sqrt{K}} + \sqrt{\Delta_\theta}\right]} - \frac{\alpha_K}{\lambda}.$$

Finally, note that implies that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T \Lambda_{\theta^t}\right] &\le \frac{1}{2\eta} + \eta\,\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{\mathcal{D}\sim\pi_{\mathtt{D}}}\|\widetilde{g}(\theta^t;\mathcal{D})\|^2\right] \\ &\le \frac{1}{2\eta} + \frac{\eta}{M}\mathbb{E}\left[\sum_{t=1}^T \Delta_{\theta^t}\right] + \frac{\sigma_\star}{\lambda}\left[2\,\mathbb{E}\left[\sum_{t=1}^T\sqrt{\Delta_\theta}\right] + \frac{T}{\sqrt{K}}\right].\end{aligned}$$

Therefore, we define $\Delta = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T \Delta_{\theta^t}\right]$, and then by Cauchy inequality,

$$\begin{aligned}\frac{1}{2T\eta} + \frac{\eta}{M}\Delta + \frac{\eta\sigma_\star}{\lambda}\left[2\sqrt{\Delta} + \frac{1}{\sqrt{K}}\right] &\ge \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T \Lambda_{\theta^t}\right] \\ &\ge \frac{1}{2T}\mathbb{E}\left[\sum_{t=1}^T \frac{\Delta_{\theta^t}^2}{2\lambda\Delta_{\theta^t} + M\sigma_\star\left[\frac{1}{\sqrt{K}} + \sqrt{\Delta_{\theta^t}}\right]}\right] - \frac{\alpha_K}{\lambda} \\ &\ge \frac{1}{2}\frac{\Delta^2}{2\lambda\Delta + M\sigma_\star\left[\frac{1}{\sqrt{K}} + \sqrt{\Delta}\right]} - \frac{\alpha_K}{\lambda}.\end{aligned}$$

Re-organizing, we know as long as $\frac{\lambda\eta}{M} \le \frac{1}{8}$, it holds that

$$\Delta \lesssim \left(\frac{M\sigma_\star}{T\eta}\right)^{2/3} + \frac{\lambda}{T\eta} + (\eta\sigma_\star)^2 + \frac{\eta M\sigma_\star^2}{\lambda} + \left(\frac{M\sigma_\star}{\lambda K}\right)^{2/3} + \frac{B}{K}.$$

In particular, we choose $\lambda = \frac{M}{8\eta}$ and require $\eta \le \frac{1}{24B}$, we have

$$\Delta \lesssim \left(\frac{M\sigma_\star}{T\eta}\right)^{2/3} + \frac{M}{T\eta^2} + (\eta\sigma_\star)^2 + \frac{B}{K}.$$

Choosing $\eta = \min\left\{\frac{1}{24B}, \left(\frac{M}{\sigma_\star^2 T}\right)^{1/4}\right\}$, we have

$$\Delta \lesssim \sqrt{\frac{\sigma_\star^2 M}{T}} + \frac{B^2 M}{T} + \frac{B}{K},$$

which implies

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T \mathtt{Cov}_N(\pi_{\theta^t})\right] \le \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T \frac{1}{M}\Delta_{\theta^t}\right] \le \sqrt{\frac{\sigma_\star^2}{TM}} + \frac{B^2}{T} + \frac{B}{KM}.$$

This is the desired upper bound. □

Remark K.1 (Comparison to standard convex optimization analyses)**.** *On a technical level, we find the proof of Theorem 5.1 to be interesting because it does not pass through KL divergence as an intermediate quantity. More broadly, we do not know how to derive the result as an application of standard analysis techniques in optimization (e.g., via a gradient dominance or PL-type condition), but it would be interesting to see if there is a connection.*[8]

K.3 Proof of Theorem 4.2

We prove the following slightly stronger result. Theorem 4.2 follows immediately by combining Theorem K.1 and Proposition D.10.

Theorem K.1. *Suppose that Assumption 2.2 holds. Then the MLE $\widehat{\pi}$ achieves*

$$\mathbb{E}_{\mathcal{D}}[D_{\mathsf{seq},N}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi})] \lesssim \sqrt{\frac{\sigma_\star^2 \log N}{n}} + \frac{B^2 \log N}{n},$$

for any parameter $N \geq 1$, where the divergence $D_{\mathsf{seq},N}(\cdot \,\|\, \cdot)$ is defined in Proposition D.10.

The following lemma follows from the optmality of the MLE $\widehat{\pi} = \pi_{\widehat{\theta}}$, i.e.,

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \widehat{\mathbb{E}}_{\mathcal{D}}[\log \pi_\theta(y_{1:H} \mid x)].$$

Lemma K.1. *Denote*

$$E_1 := \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H} \epsilon_{\widehat{\theta}}(x, y_{1:h-1})\right] = \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H} D_{\mathsf{KL}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}) \,\|\, \widehat{\pi}(\cdot \mid x, y_{1:h-1}))\right]. \tag{59}$$

Then it holds that $\mathbb{E}[E_1] \leq \frac{2\sigma_\star}{\sqrt{n}}$.

In the following, we prove concentration bounds on E_1. For simplicity, we denote $A = \log N$.

Lemma K.2. *Fix any $\Delta \in (0, \frac{1}{100B}]$, $\delta \in (0,1)$, and let $J = \exp\big(\frac{1}{\Delta^2} + 2\big)\log(1/\delta)$. Let $\Theta' := \{\theta_1, \cdots, \theta_J\}$, where $\theta_1, \cdots, \theta_J \sim \mathcal{N}(0, \Delta^2 I)$ are sampled i.i.d. Then the following holds with probability at least $1-\delta$ over the randomness of Θ' and $\mathcal{D}$:*

(1) For any $j \in [J]$, it holds that

$$\mathbb{E}_{\pi_{\mathsf{D}}} \min\left\{A, \sum_{h=1}^{H} \epsilon_{\theta_j}(x, y_{1:h-1})\right\} \leq 2\widehat{\mathbb{E}}_{\mathcal{D}} \min\left\{A, \sum_{h=1}^{H} \epsilon_{\theta_j}(x, y_{1:h-1})\right\} + \frac{8A\log(4J/\delta)}{n}.$$

(2) There exists $j \in [J]$ such that

$$\mathbb{E}_{\pi_{\mathsf{D}}} \min\left\{A, \sum_{h=1}^{H} \epsilon_{\widehat{\theta}}(x, y_{1:h-1})\right\} \leq 2\,\mathbb{E}_{\pi_{\mathsf{D}}} \min\left\{A, \sum_{h=1}^{H} \epsilon_{\theta_j}(x, y_{1:h-1})\right\} + 100\Delta^2\sigma_\star^2, \tag{60}$$

and

$$\begin{aligned}\widehat{\mathbb{E}}_{\mathcal{D}} \min\left\{A, \sum_{h=1}^{H} \epsilon_{\theta_j}(x, y_{1:h-1})\right\} \leq 2\widehat{\mathbb{E}}_{\mathcal{D}} \min\left\{A, \sum_{h=1}^{H} \epsilon_{\widehat{\theta}}(x, y_{1:h-1})\right\} \\ + 100\Delta^2 \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H} \mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\right].\end{aligned} \tag{61}$$

[8] We further note that the inherent variance $\sigma_\star^2$ corresponds to the gradient variance at the true parameter $\theta^\star$, and hence is tighter than typical analyses that depend on global notions of variance.

Now, we condition on the success event $\mathcal{E}$ of Lemma K.2, and let $j \in [J]$ be an index such that (60) and (61) hold. Then, we can upper bound (recall that $A = \log N$)

$$\begin{aligned}
D_{\mathsf{seq},N}\big(\pi_{\mathsf{D}} \,\|\, \pi_{\widehat{\theta}}\big) &= \mathbb{E}_{\pi_{\mathsf{D}}} \min\left\{A, \sum_{h=1}^{H} \epsilon_{\widehat{\theta}}(x, y_{1:h-1})\right\} \\
&\le 2\,\mathbb{E}_{\pi_{\mathsf{D}}} \min\left\{A, \sum_{h=1}^{H} \epsilon_{\theta_j}(x, y_{1:h-1})\right\} + 100\Delta^2\sigma_\star^2 \\
&\le 4\widehat{\mathbb{E}}_{\mathcal{D}} \min\left\{A, \sum_{h=1}^{H} \epsilon_{\theta_j}(x, y_{1:h-1})\right\} + \frac{16A\log(4J/\delta)}{n} + 100\Delta^2\sigma_\star^2 \\
&\le 8\widehat{\mathbb{E}}_{\mathcal{D}} \min\left\{A, \sum_{h=1}^{H} \epsilon_{\widehat{\theta}}(x, y_{1:h-1})\right\} \\
&\quad + 400\Delta^2\widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H} \mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\right] + \frac{16A\log(4J/\delta)}{n} + 100\Delta^2\sigma_\star^2.
\end{aligned}$$

where the first inequality uses (60), the second inequality uses Lemma K.2 (1), and the third inequality uses (61). Therefore, we denote $\sigma^2(\mathcal{D}) := \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H} \mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\right]$, and we have shown that for any $\delta \in (0,1)$, any $\Delta(0, \frac{1}{100B}]$, it holds that

$$\mathbb{P}_{\mathcal{D}\sim\pi_{\mathsf{D}}}\left(D_{\mathsf{seq},N}\big(\pi_{\mathsf{D}} \,\|\, \pi_{\widehat{\theta}}\big) \le C\left(E_1 + \Delta^2\sigma^2(\mathcal{D}) + \Delta^2\sigma_\star^2 + \frac{A}{n}\left(\frac{1}{\Delta^2} + \log(1/\delta)\right)\right)\right) \le \delta,$$

where $C > 0$ is an absolute constant.

By the arbitrariness of $\delta \in (0,1)$, taking expectation gives

$$\begin{aligned}
\mathbb{E}\big[D_{\mathsf{seq},N}\big(\pi_{\mathsf{D}} \,\|\, \pi_{\widehat{\theta}}\big)\big] &\le C\left(\mathbb{E}[E_1] + \Delta^2\,\mathbb{E}[\sigma^2(\mathcal{D})] + \Delta^2\sigma_\star^2 + \frac{A}{n}\left(\frac{1}{\Delta^2} + 1\right)\right) \\
&\le 2C\left(\sqrt{\frac{\sigma_\star^2}{n}} + \Delta^2\sigma_\star^2 + \frac{A}{n\Delta^2}\right).
\end{aligned}$$

Choosing $\Delta = \min\left\{\frac{1}{100B}, \left(\frac{A}{\sigma_\star^2 n}\right)^{1/4}\right\}$ completes the proof. □

K.3.1 Proofs of the supporting lemmas

Proof of Lemma K.1. Recall that $\widehat{\pi} = \pi_{\widehat{\theta}}$, where $\widehat{\theta} = \arg\max_{\theta\in\Theta} \sum_{(x,y_{1:H})\in\mathcal{D}} \log \pi_\theta(y_{1:H} \mid x)$. Then by the concavity, we know

$$\left\langle \widehat{\mathbb{E}}_{\mathcal{D}}[\log \pi_\theta(y_{1:H} \mid x)], \theta - \widehat{\theta} \right\rangle \le 0, \qquad \forall \theta \in \Theta$$

where we recall that $\widehat{\mathbb{E}}_{\mathcal{D}}$ is the empirical distribution $(x, y_{1:H}) \sim \mathsf{Unif}(\mathcal{D})$. Using the expressing (54) and $\theta^\star \in \Theta$, we know

$$\widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H} \left\langle \big(\phi(x, y_{1:h}) - \bar{\phi}_{\widehat{\theta}}(x, y_{1:h-1})\big), \theta^\star - \widehat{\theta} \right\rangle\right] \le 0.$$

Therefore, combining the inequality above with Eq. (56), we have

$$\begin{aligned}\widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\epsilon_{\widehat{\theta}}(x,y_{1:h-1})\right] &= \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}D_{\mathsf{KL}}(\pi_{\mathfrak{D}}(\cdot\mid x,y_{1:h-1})\,\|\,\widehat{\pi}(\cdot\mid x,y_{1:h-1}))\right]\\ &\le \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\left\langle\bar{\phi}_{\theta^\star}(x,y_{1:h-1})-\bar{\phi}_{\widehat{\theta}}(x,y_{1:h-1}),\theta^\star-\widehat{\theta}\right\rangle\right]\\ &\le \widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\left\langle\bar{\phi}_{\theta^\star}(x,y_{1:h-1})-\phi(x,y_{1:h}),\theta^\star-\widehat{\theta}\right\rangle\right]\\ &\le 2\left\|\widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\phi^\star(x,y_{1:h})\right]\right\| =: E_1',\end{aligned}$$

where we recall that $\phi^\star(x,y_{1:h}) := \phi(x,y_{1:h})-\bar{\phi}_{\theta^\star}(x,y_{1:h-1})$. By definition, it holds that $\mathbb{E}_{\pi_{\mathfrak{D}}}[\phi^\star(x,y_{1:h})\mid x,y_{1:h-1}]=0$, and hence

$$\begin{aligned}\mathbb{E}(E_1')^2 &= \mathbb{E}\left\|\widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\phi^\star(x,y_{1:h})\right]\right\|^2\\ &= \frac{1}{n}\,\mathbb{E}_{\pi_{\mathfrak{D}}}\left\|\sum_{h=1}^{H}\phi^\star(x,y_{1:h})\right\|^2 = \frac{1}{n}\,\mathbb{E}_{\pi_{\mathfrak{D}}}\left[\sum_{h=1}^{H}\|\phi^\star(x,y_{1:h})\|^2\right] = \frac{\sigma_\star^2}{n}.\end{aligned}$$

This gives the desired upper bound. □

Proof of Lemma K.2. By Freedman inequality (Lemma H.3) and union bound, it is clear that (1) holds with probability at least $1-\frac{\delta}{2}$. In the following, we prove (2).

Define the following weight function $\alpha=\alpha_{\widehat{\theta}}:\mathcal{X}\times\mathcal{A}^\star\to[0,1]$:

$$\alpha_{\widehat{\theta}}(x,y_{1:h-1}) = \begin{cases}1, & \sum_{j\le h-1}\epsilon_{\widehat{\theta}}(x,y_{1:j})\le A,\\ 0, & \sum_{j<h-1}\epsilon_{\widehat{\theta}}(x,y_{1:j})\ge A,\\ \frac{A-\sum_{j<h-1}\epsilon_{\widehat{\theta}}(x,y_{1:j})}{\epsilon_{\widehat{\theta}}(x,y_{1:h-1})}, & \text{otherwise.}\end{cases}$$

Then, by Lemma K.4, it holds that for any $\theta\in\Theta$,

$$\begin{aligned}\mathbb{E}_{\pi_{\mathfrak{D}}}\min\left\{A,\sum_{h=1}^{H}\epsilon_{\widehat{\theta}}(x,y_{1:h-1})\right\} &\le 2\,\mathbb{E}_{\pi_{\mathfrak{D}}}\min\left\{A,\sum_{h=1}^{H}\epsilon_{\theta}(x,y_{1:h-1})\right\}\\ &\quad+2\,\mathbb{E}_{\pi_{\mathfrak{D}}}\left[\sum_{h=1}^{H}\alpha(x,y_{1:h-1})\mathsf{F}\big(\epsilon_{\widehat{\theta}}(x,y_{1:h-1}),\epsilon_{\theta}(x,y_{1:h-1})\big)\right],\end{aligned}$$

and

$$\begin{aligned}\widehat{\mathbb{E}}_{\mathcal{D}}\min\left\{A,\sum_{h=1}^{H}\epsilon_{\theta}(x,y_{1:h-1})\right\} &\le 2\widehat{\mathbb{E}}_{\mathcal{D}}\min\left\{A,\sum_{h=1}^{H}\epsilon_{\widehat{\theta}}(x,y_{1:h-1})\right\}\\ &\quad+\widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H}\alpha(x,y_{1:h-1})\mathsf{F}\big(\epsilon_{\widehat{\theta}}(x,y_{1:h-1}),\epsilon_{\theta}(x,y_{1:h-1})\big)\right],\end{aligned}$$

Therefore, it remains to control the error $\sum_{h=1}^{H}\alpha(x,y_{1:h-1})\mathsf{F}\big(\epsilon_{\widehat{\theta}}(x,y_{1:h-1}),\epsilon_{\theta}(x,y_{1:h-1})\big)$ under both $\mathbb{E}_{\pi_{\mathfrak{D}}}[\cdot]$ and $\widehat{\mathbb{E}}_{\mathcal{D}}[\cdot]$. We prove the following lemma, which leverages the structure of Gaussian distribution.

Lemma K.3. *For any $K\ge 1$, $\Delta\in(0,\frac{1}{40KB}]$, $\theta\in\mathbb{B}_2(1)$, distributions $\mu_1,\cdots,\mu_K$ over $\mathcal{Z}:=\mathcal{X}\times\mathcal{A}^\star$, and weight function $\alpha:\mathcal{Z}\to[0,1]$, it holds that*

$$-\log\mathbb{P}_{\theta'\sim\mathcal{N}(0,\Delta^2)}\big(\forall i\in[K],\mathbb{E}_{z\sim\mu_i}\alpha(z)\mathsf{F}(\epsilon_\theta(z),\epsilon_{\theta'}(z))\le 22K^2\Delta^2\,\mathbb{E}_{z\sim\mu_i}\mathrm{Var}_{\pi_{\mathfrak{D}}}(z)\big)\le\frac{1}{\Delta^2}+2.$$

In the following, we apply Lemma K.3 with $K = 2$, parameter $\theta = \widehat{\theta}$, weight function α, and the distributions μ_1, μ_2 defined as follows:

- Let μ_1 be the distribution of $x' = (x, y_{1:h-1})$ under $x \sim \mu$, $y_{1:H} \sim \pi_{\mathsf{D}}(\cdot \mid x)$ and $h \sim \mathsf{Unif}([H])$.
- Let μ_2 be the distribution of $x' = (x^t, y^t_{1:h-1})$ under $t \sim \mathsf{Unif}([n])$ and $h \sim \mathsf{Unif}([H])$.

By definition, it holds that

$$\mathbb{E}_{z\sim\mu_1} \alpha(z)\mathsf{F}(\epsilon_\theta(z), \epsilon_{\theta'}(z)) = \frac{1}{H}\mathbb{E}_{\pi_{\mathsf{D}}}\left[\sum_{h=1}^{H} \alpha(x, y_{1:h-1})\mathsf{F}\big(\epsilon_{\widehat{\theta}}(x, y_{1:h-1}), \epsilon_\theta(x, y_{1:h-1})\big)\right],$$

$$\mathbb{E}_{z\sim\mu_1} \mathrm{Var}_{\pi_{\mathsf{D}}}(z) = \frac{1}{H}\mathbb{E}_{\pi_{\mathsf{D}}}\left[\sum_{h=1}^{H} \mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\right] = \frac{\sigma_\star^2}{H},$$

$$\mathbb{E}_{z\sim\mu_2} \alpha(z)\mathsf{F}(\epsilon_\theta(z), \epsilon_{\theta'}(z)) = \frac{1}{H}\widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H} \alpha(x, y_{1:h-1})\mathsf{F}\big(\epsilon_{\widehat{\theta}}(x, y_{1:h-1}), \epsilon_\theta(x, y_{1:h-1})\big)\right],$$

$$\mathbb{E}_{z\sim\mu_2} \mathrm{Var}_{\pi_{\mathsf{D}}}(z) = \frac{1}{H}\widehat{\mathbb{E}}_{\mathcal{D}}\left[\sum_{h=1}^{H} \mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\right].$$

Then, we consider the following set

$$\Theta^+_{\widehat{\theta}} := \big\{\forall i \in \{1, 2\}, \mathbb{E}_{z\sim\mu_i} \alpha(z)\mathsf{F}(\epsilon_\theta(z), \epsilon_{\theta'}(z)) \le 100\Delta^2\, \mathbb{E}_{z\sim\mu_i} \mathrm{Var}_{\pi_{\mathsf{D}}}(z)\big\}.$$

By Lemma K.3, it holds that

$$q_{\widehat{\theta}} := \mathbb{P}_{\theta'\sim\mathcal{N}(0,\Delta^2 I)}(\theta' \in \Theta^+_{\widehat{\theta}}) \ge \exp\left(-\frac{1}{\Delta^2} - 2\right).$$

Therefore, we have

$$\begin{aligned}\mathbb{P}\Big(\forall j \in [N], \theta_j \notin \Theta^+_{\widehat{\theta}} \mid \widehat{\theta}\Big) &= \mathbb{P}_{\theta_1,\cdots,\theta_J\sim\mathcal{N}(0,\Delta^2 I)}\Big(\forall j \in [N], \theta_j \notin \Theta^+_{\widehat{\theta}}\Big) \\ &\le (1 - q_{\widehat{\theta}})^N \le \exp\big(-N q_{\widehat{\theta}}\big) \le \frac{\delta}{2},\end{aligned}$$

and hence $\mathbb{P}\Big(\exists j \in [N], \theta_j \in \Theta^+_{\widehat{\theta}}\Big) \ge 1 - \frac{\delta}{2}$. The proof of (2) is hence completed. □

Proof of Lemma K.3. By definition, we have $\pi_{\theta'}(y \mid z) \propto_y \pi_\theta(y \mid z) \cdot \exp(\langle\theta' - \theta, \phi(z, y)\rangle)$, i.e.,

$$\log \pi_{\theta'}(y \mid z) - \log \pi_\theta(y \mid z) = \langle\theta' - \theta, \phi(z, y)\rangle - \log \mathbb{E}_{y\sim\pi_\theta(\cdot|z)} \exp(\langle\theta' - \theta, \phi(z, y)\rangle).$$

Therefore,

$$\begin{aligned}\epsilon_\theta(z) - \epsilon_{\theta'}(z) &= D_{\mathsf{KL}}(\pi_{\mathsf{D}}(\cdot \mid z) \,\|\, \pi_\theta(\cdot \mid z)) - D_{\mathsf{KL}}(\pi_{\mathsf{D}}(\cdot \mid z) \,\|\, \pi_{\theta'}(\cdot \mid z)) \\ &= \mathbb{E}_{\pi_{\mathsf{D}}(\cdot|z)}\langle\theta' - \theta, \phi(z, y)\rangle - \log \mathbb{E}_{y\sim\pi_\theta(\cdot|z)} \exp(\langle\theta' - \theta, \phi(z, y)\rangle) \\ &= \langle\theta' - \theta, \bar{\phi}_{\theta^\star}(z) - \bar{\phi}_\theta(z)\rangle - \log \mathbb{E}_{y\sim\pi_\theta(\cdot|z)} \exp\big(\langle\theta' - \theta, \phi(z, y) - \bar{\phi}_\theta(z)\rangle\big),\end{aligned}$$

where we recall that $\bar{\phi}_\theta(z) = \mathbb{E}_{y\sim\pi_\theta(\cdot|z)}[\phi(z, y)]$.

In the following, we denote $\phi_\theta(z, y) := \phi(z, y) - \bar{\phi}_\theta(z)$, and

$$\begin{aligned}E^+_{\theta'}(z) &:= \log \mathbb{E}_{y\sim\pi_\theta(\cdot|z)} \exp(\langle\theta' - \theta, \phi_\theta(z, y)\rangle), \\ E^-_{\theta'}(z) &:= \langle\theta' - \theta, \bar{\phi}_{\theta^\star}(z) - \bar{\phi}_\theta(z)\rangle.\end{aligned}$$

We first bound $E^+_{\theta'}(z)$. By definition, we have $E^+_{\theta'}(z) = D_{\mathsf{KL}}(\pi_\theta(\cdot \mid z) \,\|\, \pi_{\theta'}(\cdot \mid z)) \ge 0$. Further, using Jensen's inequality, for any $z \in \mathcal{Z}$, we have

$$\begin{aligned}\mathbb{E}_{\theta'\sim\mathcal{N}(\theta,\Delta^2 I)}\big[E^+_{\theta'}(z)\big] &\le \log \mathbb{E}_{\theta'\sim\mathcal{N}(\theta,\Delta^2 I)}\, \mathbb{E}_{y\sim\pi_\theta(\cdot|z)}[\exp(\langle\theta' - \theta, \phi_\theta(z, y)\rangle)] \\ &= \log \mathbb{E}_{y\sim\pi_\theta(\cdot|z)} \exp\left(\frac{1}{2}\Delta^2\|\phi_\theta(z, y)\|^2\right) \\ &\le \Delta^2\, \mathbb{E}_{y\sim\pi_\theta(\cdot|z)}\|\phi_\theta(z, y)\|^2,\end{aligned}$$

where the last inequality follows from $e^t \le 1 + 2t$ for $t \in [0, 1]$. Further, using Lemma H.5, we have

$$\begin{aligned}\mathbb{E}_{y\sim\pi_\theta(\cdot|z)}\|\phi_\theta(z,y)\|^2 &= \mathbb{E}_{y\sim\pi_\theta(\cdot|z)}\|\phi(z,y) - \phi_\theta(z)\|^2 \\ &\le 3\,\mathbb{E}_{y\sim\pi_{\mathsf{D}}(\cdot|z)}\|\phi(z,y) - \phi_{\theta^\star}(z)\|^2 + 4B^2 D_{\mathsf{KL}}(\pi_{\mathsf{D}}(\cdot \mid z)\,\|\,\pi_\theta(\cdot \mid z)) \\ &= 3\mathrm{Var}_{\pi_{\mathsf{D}}}(z) + 4B^2\epsilon_\theta(z).\end{aligned}$$

Next, we bound $|E^-_{\theta'}(z)|$. Under $\theta' \sim \mathcal{N}(\theta, \Delta^2 I)$, it is clear that $\langle \theta' - \theta, \bar\phi_{\theta^\star}(z) - \bar\phi_\theta(z)\rangle \sim \mathcal{N}(0, \Delta^2\|\bar\phi_{\theta^\star}(z) - \bar\phi_\theta(z)\|^2)$ for any fixed z. Therefore, it holds that

$$\begin{aligned}\mathbb{E}_{\theta'\sim\mathcal{N}(\theta,\Delta^2 I)}\big|E^-_{\theta'}(z)\big| &= \sqrt{\frac{2}{\pi}}\Delta\cdot\|\bar\phi_{\theta^\star}(z) - \bar\phi_\theta(z)\| \\ &\le \Delta\cdot\Big(2\sqrt{\mathrm{Var}_{\pi_{\mathsf{D}}}(z)\cdot\epsilon_\theta(z)} + 3B\epsilon_\theta(z)\Big) \\ &\le \Big(\frac{1}{8K} + 3B\Delta\Big)\epsilon_\theta(z) + 8K\Delta^2\mathrm{Var}_{\pi_{\mathsf{D}}}(z).\end{aligned}$$

where the second line uses (56).

Combining the inequalities above, we know that for $i \in [K]$, it holds that

$$\begin{aligned}&\mathbb{E}_{\theta'\sim\mathcal{N}(\theta,\Delta^2 I)}\big[\mathbb{E}_{z\sim\mu_i}\big[\alpha(z)E^+_{\theta'}(z)\big]\big] \le \Delta^2\,\mathbb{E}_{z\sim\mu_i}\big[3\mathrm{Var}_{\pi_{\mathsf{D}}}(z) + 4B^2\alpha(z)\epsilon_\theta(z)\big], \\ &\mathbb{E}_{\theta'\sim\mathcal{N}(\theta,\Delta^2 I)}\big[\mathbb{E}_{z\sim\mu_i}\big[\alpha(z)\big|E^-_{\theta'}(z)\big|\big]\big] \le \mathbb{E}_{z\sim\mu_i}\left[8K\Delta^2\mathrm{Var}_{\pi_{\mathsf{D}}}(z) + \Big(\frac{1}{8K} + 3B\Delta\Big)\alpha(z)\epsilon_\theta(z)\right],\end{aligned}$$

and hence by Markov's inequality, it holds that $p := \mathbb{P}_{\theta'\sim\mathcal{N}(\theta,\Delta^2 I)}(\theta' \notin \Theta^-) \ge \frac{1}{2}$, where we denote $\Theta^- = \cup_{i\in[K]}\Theta_i$, and

$$\Theta_i := \left\{\theta' \in \mathbb{R}^d : \mathbb{E}_{z\sim\mu_i}\,\alpha(z)|\epsilon_\theta(z) - \epsilon_{\theta'}(z)| \ge \mathbb{E}_{z\sim\mu_i}\left[(6K + 16K^2)\Delta^2\mathrm{Var}_{\pi_{\mathsf{D}}}(z) + \frac{1}{2}\alpha(z)\epsilon_\theta(z)\right]\right\}.$$

Note that $D_{\mathsf{KL}}\big(\mathcal{N}(\theta, \Delta^2 I)\,\|\,\mathcal{N}(0, \Delta^2 I)\big) = \frac{\|\theta\|^2}{2\Delta^2} \le \frac{1}{2\Delta^2}$. Hence, by data-processing inequality, we can bound $q := \mathbb{P}_{\theta'\sim\mathcal{N}(0,\Delta^2 I)}(\theta' \notin \Theta^-)$ as

$$\begin{aligned}\frac{1}{2\Delta^2} &\ge D_{\mathsf{KL}}\big(\mathcal{N}(\theta, \Delta^2 I)\,\|\,\mathcal{N}(0, \Delta^2 I)\big) \ge D_{\mathsf{KL}}(\mathrm{Ber}(p)\,\|\,\mathrm{Ber}(q)) \\ &= p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \ge \frac{1}{2}\log(1/q) - \log 2,\end{aligned}$$

impling that $-\log q \le \frac{1}{\Delta^2} + 2$. This is the desired result. □

Lemma K.4. *Suppose that $a_1, \cdots, a_H, b_1, \cdots, b_H \ge 0$. Let*

$$\alpha_h = \begin{cases} 1, & \sum_{j\le h} a_j \le A, \\ 0, & \sum_{j<h} a_j > A, \\ \frac{A - \sum_{j<h} a_j}{a_h}, & \textit{otherwise}. \end{cases}$$

Then clearly $\alpha_h \in [0, 1]\ \forall h \in [H]$, and it holds that $\sum_{h=1}^H \alpha_h a_h = \min\Big\{A, \sum_{h=1}^H a_h\Big\}$, and

$$\begin{aligned}\min\left\{A, \sum_{h=1}^H a_h\right\} &\le 2\min\left\{A, \sum_{h=1}^H b_h\right\} + 2\sum_{h=1}^H \alpha_h \mathsf{F}(a_h, b_h), \\ \min\left\{A, \sum_{h=1}^H b_h\right\} &\le 2\min\left\{A, \sum_{h=1}^H a_h\right\} + \sum_{h=1}^H \alpha_h \mathsf{F}(a_h, b_h),\end{aligned}$$

where we recall that $\mathsf{F}(a, b) = |a - b| - \frac{1}{2}a$.

Proof of Lemma K.4. Fix the sequence $a_1, \cdots, a_H$.

We first prove $\sum_{h=1}^H \alpha_h a_h = \min\Big\{A, \sum_{h=1}^H a_h\Big\}$.

Case 1: $\sum_{h=1}^H a_h \le A$**.** In this case, $\alpha_h = 1 \forall h \in [H]$, and the equation holds trivially.

Case 2: $\sum_{h=1}^H a_h > A$**.** In this case, we let $\ell \in [H]$ be the maximal index such that $\alpha_\ell > 0$. Then, by definition, $\sum_{j<\ell} a_j \le A$ and $\sum_{j\le\ell} a_j > A$, and $\alpha_\ell = \frac{A - \sum_{j<\ell} a_j}{a_\ell}$. Hence,

$$\sum_{h=1}^H \alpha_h a_h = \sum_{h=1}^\ell \alpha_h a_h = \sum_{j<\ell} a_j + \alpha_\ell a_\ell = A.$$

We note that from the proof above, we also know that for any sequence $(c_1, \cdots, c_H)$ such that $c_h \ge a_h$ for $h \in [H]$, we have $\min\Big\{A, \sum_{h=1}^H c_h\Big\} \le \sum_{h=1}^H \alpha_h c_h$.

Next, we prove the inequalities. We note that

$$\sum_{h=1}^H \alpha_h \mathsf{F}(a_h, b_h) = \sum_{h=1}^H \alpha_h |a_h - b_h| - \frac{1}{2}\sum_{h=1}^H \alpha_h a_h,$$

or equivalently,

$$\sum_{h=1}^H \alpha_h |a_h - b_h| = \sum_{h=1}^H \alpha_h \mathsf{F}(a_h, b_h) + \frac{1}{2}\min\left\{A, \sum_{h=1}^H a_h\right\}.$$

Therefore,

$$\begin{aligned}\min\left\{A, \sum_{h=1}^H a_h\right\} &= \sum_{h=1}^H \alpha_h a_h \le \min\left\{A, \sum_{h=1}^H b_h\right\} + \sum_{h=1}^H \alpha_h |a_h - b_h| \\ &= \min\left\{A, \sum_{h=1}^H b_h\right\} + \sum_{h=1}^H \alpha_h \mathsf{F}(a_h, b_h) + \frac{1}{2}\min\left\{A, \sum_{h=1}^H a_h\right\}.\end{aligned}$$

Re-organizing yields the first inequality. Similarly, we have

$$\begin{aligned}\min\left\{A, \sum_{h=1}^H b_h\right\} &\le \min\left\{A, \sum_{h=1}^H (a_h + |a_h - b_h|)\right\} \le \sum_{h=1}^H \alpha_h (a_h + |a_h - b_h|) \\ &= \frac{3}{2}\min\left\{A, \sum_{h=1}^H a_h\right\} + \sum_{h=1}^H \alpha_h \mathsf{F}(a_h, b_h).\end{aligned}$$

The proof is hence completed. □

K.4 Proof of Proposition 5.1 (Lower Bound)

In the following, we construct $\mathcal{X} = \mathbb{R} \sqcup \{-, +\}$, $\mathcal{Y} = \{-1, 0, 1\}$ and $\Theta = \mathbb{B}_2(1)$ with $d = 2$. The feature map ϕ satisfies $\phi(x, y_{1:h}) = \phi(x, y_h)$, i.e., $y_{1:H} \sim \pi_\theta(\cdot \mid x)$ are i.i.d. We fix $B \ge c_B \log H$ for a sufficiently large constant $c_B > 0$.

Case 1: $\eta \ge \frac{8}{HB}$**.** We define $\bar\eta := \eta \cdot HB$ and $\alpha = \frac{\bar\eta}{2(\bar\eta - 1)} \le \frac{5}{8}$. Let

$$v_0 = [1; 0], \qquad v_1 = [\alpha; \sqrt{1 - \alpha^2}], \qquad v_{-1} = [\alpha; -\sqrt{1 - \alpha^2}].$$

For $y_{1:h} \in \mathcal{A}^h$, we define $\phi(\eta, y_{1:h}) = B v_{y_h}$, and we consider the problem instance with μ supported on $x = \eta$ (i.e., $\mu(\eta) = 1$) and $\theta^\star = v_0$.

In the following, we omit the dependence on $x = \eta$. Then, under this construction, we have

$$\pi_\theta(y_h \mid y_{1:h}) = \frac{\exp(B\langle\theta, v_{y_h}\rangle)}{\sum_{y' \in \mathcal{Y}} \exp(B\langle\theta, v_{y'}\rangle)} = \pi_\theta(y_h).$$

We study the SGD update starting from $\theta^0 = v_1$. By definition,

$$\nabla \log \pi_\theta(y_{1:H}) = \sum_{h=1}^{H} \Big(\phi(y_h) - \mathop{\mathbb{E}}_{y\sim\pi_\theta}[\phi(y)] \Big).$$

In the following, we denote $\widehat{F}(y_{1:H}) := \frac{1}{H}\sum_{h=1}^{H} v_{y_h}$, and

$$F(\theta) := \mathop{\mathbb{E}}_{y\sim\pi_\theta}[\phi(y)] = \frac{\sum_{y\in\mathcal{Y}} y \exp(B\langle\theta, v_y\rangle)}{\sum_{y'\in\mathcal{Y}} \exp(B\langle\theta, v_{y'}\rangle)}$$

Then, the SGD update can be written as

$$\theta^{t+1} = \mathrm{Proj}_\Theta\Big(\theta^t + \bar\eta\Big(\widehat{F}(y^t_{1:H}) - F(\theta^t)\Big)\Big).$$

We make the following claims.

Claim 1. For $y \in \mathcal{Y}$ and $\|\theta - v_y\| \le \frac{1}{16}$, it holds that $1 - \pi_\theta(y) \le 2e^{-B/4} =: \epsilon_1$ and hence $\|F(\theta) - v_y\| \le 2\epsilon_1$.

Claim 2. Suppose that $\epsilon_1 \le \min\{\frac{1}{4nH}, \frac{1}{5HB^2}\}$. Then it holds that $\mathrm{Var}_{\pi_{\mathtt{D}}}[\phi(y_1)] \le \frac{1}{H}$ and $\sigma_\star \le 1$, and $\mathtt{Cov}_N(\pi_{\mathtt{D}} \,\|\, \pi_\theta) \ge 1 - \frac{1}{2n}$ for $\log N \le \frac{HB}{8}$ and $\theta \in \Theta$ such that $\min\{\|\theta - v_1\|, \|\theta - v_{-1}\|\} \le \frac{1}{16}$. Further, with probability at least 0.5, it holds that $\widehat{F}(y^t_{1:H}) = e_0$ for all $t \in [n]$.

In the following, we condition on this event.

Claim 3. By definition, for $y \in \{-1, 1\}$, we have $\|v_y + \bar\eta(v_0 - v_y)\| = \bar\eta - 1$ and $v_{-1} + v_1 = \frac{\bar\eta}{\bar\eta - 1} v_0$.

Claim 4. Let $\epsilon = 16\epsilon_1 + 4\epsilon_0$. Suppose that $\epsilon \le \frac{1}{16}$. Then if $\|\theta^t - v_y\| \le \epsilon$, then it holds that $\|\theta^{t+1} - v_{-y}\| \le \epsilon$.

Combining the above claims, we know that there is a constant C such that as long as $B \ge C\log(nH)$, it holds that $\sigma_\star \le 1$ and with probability at least 0.5, $\|\theta^t - v_{t(\mod 2)}\| \le \frac{1}{16}$ for all $t \in [n]$. Therefore, by Claim 2, this gives $\mathtt{Cov}_N(\pi_{\theta^t}) \ge \frac{1}{2}$ as long as $\log N \le \frac{HB}{8}$. □

Proof of the claims. To prove Claim 1, we note that $\langle\theta, v_y\rangle \ge 1 - \|\theta - v_y\| \ge \frac{15}{16}$ and for $y' \ne y$, $\langle\theta, v_{y'}\rangle \le \langle v_y, v_{y'}\rangle + \|\theta - v_y\| \le \alpha + \frac{1}{16} \le \frac{11}{16}$. Therefore,

$$1 - \pi_\theta(y) \le \frac{\sum_{y'\ne y} e^{B\langle\theta, v_{y'}\rangle}}{e^{B\langle\theta, v_y\rangle}} \le \frac{2}{e^{B/4}} = \epsilon_1.$$

In particular, we know $1 - \pi_{\mathtt{D}}(0) \le \epsilon_1$, and hence $\mathrm{Var}_{\pi_{\mathtt{D}}}[\phi(y_1)] \le 5B^2\epsilon_1$. Further, we also know $\mathbb{P}_{\pi_{\mathtt{D}}}(y_h = 0 \forall h \in [H]) \ge (1-\epsilon_1)^H \ge 1 - H\epsilon_1$. Therefore, taking the union bound, we know $\mathbb{P}(y^t_h = 0 \forall h \in [H], t \in [n]) \ge 1 - nH\epsilon_1 \ge \frac{1}{2}$.

Furthermore, for any θ such that $\min\{\|\theta - v_1\|, \|\theta - v_{-1}\|\} \le \frac{1}{16}$, as long as $\log N \le H(\log(1 - \epsilon_1) - \log(\epsilon_1))$, we have

$$\mathtt{Cov}_N(\pi_{\mathtt{D}} \,\|\, \pi_\theta) \ge \left(1 - \frac{1}{2n}\right)\mathbb{I}\{H\log\pi_{\mathtt{D}}(0) - H\log\pi_\theta(0) \ge \log N\} \ge 1 - \frac{1}{2n}.$$

In particular, this is ensured when $\log N \le \frac{HB}{8}$. This completes the proof of Claim 2.

Claim 3 follows immediately from the definition of α, v_0, v_1 and v_{-1}. Finally, we prove claim 4. We define $u^t := \theta^t + \bar\eta\Big(\widehat{F}(y^t_{1:H}) - F(\theta^t)\Big)$. Then it holds that

$$\begin{aligned}\|u^t - (\bar\eta - 1)v_{-y}\| = \|u^t - \bar\eta v_0 + (\bar\eta - 1)v_y\| &\le \|\theta^t - v_y\| + \bar\eta\|\widehat{F}(y^t_{1:H}) - v_0\| + \bar\eta\|F(\theta^t) - v_y\| \\ &\le \epsilon + \bar\eta(2\epsilon_1 + \epsilon_0) =: \epsilon'.\end{aligned}$$

In particular, it holds that $|\|u^t\| - (\bar{\eta}-1)| \le \epsilon'$ and hence $\|u^t\| \ge \bar{\eta} - 1 - \epsilon' \ge \bar{\eta} - 2 \ge 1$. Therefore, $\theta^{t+1} = \mathrm{Proj}_\Theta(u^t) = \frac{u^t}{\|u^t\|}$, and we can bound

$$\begin{aligned}\|\theta^{t+1} - v_{-y}\| &= \left\| \frac{u^t - (\bar{\eta}-1)v_{-y}}{\|u^t\|} + v_{-y}\left(\frac{\bar{\eta}-1}{\|u^t\|} - 1\right)\right\| \\ &\le \frac{\|u^t - (\bar{\eta}-1)v_{-y}\|}{\|u^t\|} + \frac{|\bar{\eta} - 1 - \|u^t\||}{\|u^t\|} \\ &\le \frac{2\epsilon'}{\|u^t\|} \le \frac{4\epsilon'}{\bar{\eta}} = \frac{4}{\bar{\eta}}\epsilon + 8\epsilon_1 + 4\epsilon_0 \le \epsilon.\end{aligned}$$

□

Case 2: $\eta \le \frac{8}{HB}$**.** Let $\overline{B} \le B$ be a parameter such that $\overline{B} \ge c_B \log(c_B nH)$ for a sufficiently large constant c_B, and we again denote $\bar{\eta} = HB\eta$.

In this case, we choose the distribution μ to be $\mu(+) = 1 - \mu(-) = \min\left\{1, \frac{BH}{512en\overline{B}^2 \log N}\right\}$, the feature map be specified as $\phi(-,\cdot) = 0$ and $\phi(+, y_{1:h}) = [y_h\overline{B}; 0]$ for $y \in \mathcal{Y}$. We choose $\theta^\star = [1; 0]$. Note that $\pi_{\mathtt{D}}(1 \mid +) = \frac{e^{\overline{B}}}{e^{-\overline{B}}+1+e^{\overline{B}}}$, and hence $1 - \pi_{\mathtt{D}}(y_1 = 1 \mid +) \le 2e^{-\overline{B}}$. Therefore, similar to Case 1, we have the following claims.

Claim 1. It holds that $\sigma_\star \le 1$, and with probability at least 0.5, it holds that $\sum_{t=1}^n \mathbb{I}\{x^t = +\} \le 4\mu(1)n$, and for any t such that $x^t = +$, we have $y_h^t = 1$ for all $h \in [H]$.

In the following, we condition on this event.

Claim 2. For any $\theta \in \Theta$, it holds that $1 - \pi_\theta(1 \mid +) \le \frac{2}{e^{\theta[1]\overline{B}}}$, and hence when $x^t = +$, we have

$$\|\nabla \log \pi_\theta(y^t \mid x^t)\| = \|H(\overline{B} - \mathbb{E}_{y_1 \sim \pi_\theta(\cdot|+)}[\phi(y_1)])\| \le 2H\overline{B}|1 - \pi_\theta(1 \mid +)| \le \frac{4H\overline{B}}{e^{\theta[1]\overline{B}}}.$$

Note that when $x^t = -$, we have $\nabla \log \pi_\theta(y^t \mid x^t) = 0$. Therefore, it holds that

$$0 \le \theta^{t+1}[1] - \theta^t[1] \le \mathbb{I}\{x^t = +\} \cdot \frac{4\bar{\eta}\overline{B}}{Be^{\theta^t[1]\overline{B}}}.$$

Claim 3. Suppose that $e^{\theta[1]\overline{B}} \le \frac{H}{4\log N}$. Then it holds that $\mathtt{Cov}_N(\pi_\theta) \ge \frac{1}{2}$.

To complete the proof, we now choose $\theta' \in [-1, 1]$ such that $e^{\theta'\overline{B}} = \frac{H}{4\log N}$, and we let $\theta^0 = [\theta' - \frac{1}{\overline{B}}; 0]$. Then, using Claim 2, we know that for any $t \in [n]$, it holds that

$$\theta^t[1] - \theta^0[1] \le \sum_{t=1}^n \mathbb{I}\{x^t = +\} \cdot \frac{4\eta HB}{e^{\theta^0[1]B}} \le n \cdot \mu(+)\frac{16e\bar{\eta}\overline{B}}{Be^{\theta'B}} \le \mu(+) \cdot \frac{512e\overline{B}n\log N}{BH} \le \frac{1}{\overline{B}}.$$

Therefore, we have $\theta^t[1] \le \theta'$ and hence $\mathtt{Cov}_N(\pi_{\theta^t}) \ge \frac{\mu(+)}{2}$ for any $t \in [n]$.

It remains to prove Claim 3. We note that similar to Claim 2, $\mathbb{P}_{\pi_{\mathtt{D}}}(y_h = 1 \forall h \in [H] \mid x = +) \ge \frac{1}{2}$, and hence

$$\mathtt{Cov}_N(\pi_\theta) \ge \frac{\mu(+)}{2} \cdot \mathbb{I}\{H(\log \pi_{\mathtt{D}}(y_1 = 1 \mid +) - \log \pi_\theta(y_1 = 1 \mid +)) \ge \log N\} \ge \frac{\mu(+)}{2},$$

where we use $\log \pi_{\mathtt{D}}(y_1 = 1 \mid +) \ge \log(1 - 2e^{-B}) \ge -3e^{-B}$ and $\log \pi_\theta(y_1 = 1 \mid +) \le -\frac{1}{3e^{\theta[1]B}}$. □

K.5 Proof of the supporting results

We generalize Proposition 3.2 to show that in the worst case (where $\sigma_\star^2 \asymp HB^2$), the scaling $\mathtt{Cov}_N(\widehat{\pi}) = \Omega(\frac{H}{n\log N})$ can be unavoidable for autoregressive linear model. This implies that the dependence on $\sigma_\star^2$ is generally necessary to achieve upper bounds that do not explicitly scale with H.

Proposition K.1. *Let $H, B, N, n \geq 1$, and assume $\log N \leq c\min\{H, B^2\}$ for a sufficiently small constant $c > 0$. There exists an instance of H-dimensional autoregressive linear model class Π with $\phi : \mathcal{X} \times \mathcal{A}^\star \to \mathbb{B}_2(B)$ and $\Theta = \mathbb{B}_2(1)$, such that for any proper algorithm $\mathtt{Alg}$ with output $\widehat{\pi} = \pi_{\widehat{\theta}}$, there exists $\pi_{\mathsf{D}} \in \Pi$, such that under π_{D}, it holds that*

$$\mathbb{E}^{\pi_{\mathsf{D}},\mathtt{Alg}}[\mathsf{Cov}_N(\pi_{\mathsf{D}} \,\|\, \widehat{\pi})] \geq c \cdot \min\left\{1, \frac{H}{n \cdot \log N}\right\}.$$

Proof. We consider $\mathcal{X} = \{+, -\}$, $\mathcal{A} = \{0, 1\}$, and the distribution μ be given by $\mu(+) = 1-\mu(-) = p$, where $p \in [0, 1]$ is a pre-specified parameter. Let the feature map ϕ be given by $\phi(y_{1:h} \mid -) = 0$, $\phi(y_{1:h} \mid +) = By_h e_h$, where $(e_1, \cdots, e_H)$ is a fixed orthonormal basis of $\mathbb{R}^H$. Note that with this construction, we have $\pi_\theta(y_h = \cdot \mid -, y_{1:h-1}) = \mathrm{Ber}(1/2)$, and

$$\pi_\theta(y_h = \cdot \mid +, y_{1:h-1}) = \mathrm{Ber}\left(\frac{e^{B\theta_h}}{1 + e^{B\theta_h}}\right) =: \pi_{\theta,h}.$$

Note that for any $h \in [H]$, we can bound

$$C_0 B|\theta_h - \theta'_h| \leq D_{\mathsf{H}}(\pi_{\theta,h}, \pi_{\theta',h}) \leq C_1 B|\theta_h - \theta'_h|,$$

as long as $\theta_h \in [-\frac{1}{B}, \frac{1}{B}]$.

We fix $\epsilon \in [0, \frac{1}{\max\{\sqrt{H}, B\}}]$ to be determined later, and for any $v \in \{-1, 1\}^H$, we let $\theta_v := \epsilon \sum_{h=1}^H v_h e_h$, and

$$\Theta_0 := \left\{\theta_v : v \in \{-1, 1\}^H\right\} \subset \mathbb{B}_2(1), \qquad \Pi_0 := \{\pi_\theta : \theta \in \Theta_0\}.$$

Then a direct argument shows that when $pn \leq \frac{c_0}{B^2\epsilon^2}$ for a sufficiently small constant c_0, there exists $\theta^\star \in \Theta_0$ such that under $\pi_{\mathsf{D}} = \pi_{\theta^\star}$, it holds that

$$\sum_{h=1}^H \mathbb{P}^{\pi_{\mathsf{D}},\mathtt{Alg}}\left(|\widehat{\theta}_h - \theta^\star_h| \geq \epsilon\right) \geq cH.$$

Therefore, with probability at least $\frac{c}{2}$, it holds that $\sum_{h=1}^H \mathbb{I}\left\{|\widehat{\theta}_h - \theta^\star_h| \geq \epsilon\right\} \geq \frac{cH}{2}$, and this in turn implies

$$\sum_{h=1}^H D^2_{\mathsf{H}}\left(\pi_{\theta^\star,h}, \pi_{\widehat{\theta},h}\right) \geq c_1 H B^2 \epsilon^2.$$

Then, by Proposition D.11, we know that under the above event, as long as $\log N \leq \frac{c_1 H B^2 \epsilon^2}{2}$, we have $\mathsf{Cov}_N(\widehat{\pi}) \geq \frac{p}{2}$. Choosing $\epsilon = \sqrt{\frac{4\log N}{c_1 H B^2}}$ and $p = \min\{1, \frac{c_0}{nB^2\epsilon^2}\}$ gives the desired lower bound. □

L Proofs from Section 6

L.1 Proof of Theorem 6.1

Recall (from Eq. (11)) that we consider the token-level SGD iterates defined as

$$\theta^{t,h+1} = \mathrm{Proj}_\Theta\big(\theta^{t,h} + \eta \nabla \log \pi_{\theta^{t,h}}(y^t_h \mid x^t, y^t_{1:h-1})\big), \quad \text{for } h = 0, \cdots, H-1, \tag{62}$$

and $\theta^{t+1} \equiv \theta^{t+1,0} := \theta^{t,H}$ for $t \in [T]$, where $(x^t, y^t_{1:H}) \sim \pi_{\mathsf{D}}$.

To define the guarantee on θ^t which we are able to derive, we next define the following *test-time parameter update* $\vartheta^{\mathsf{TTT}}(x, y_{1:h}; \theta)$, for a parameter θ and prompt x. It is defined recursively for $h = 0, 1, \cdots, H-1$:

$$\vartheta^{\mathsf{TTT}}(x, y_{1:h}; \theta) := \mathrm{Proj}_\Theta\big(\vartheta^{\mathsf{TTT}}(x, y_{1:h-1}; \theta) + \eta \nabla \log \pi_{\vartheta^{\mathsf{TTT}}(x, y_{1:h-1}; \theta)}(y_h \mid x, y_{1:h-1})\big). \tag{63}$$

We then define a distribution $\pi^{\mathsf{TTT}}_\theta : \mathcal{X} \to \Delta(\mathcal{Y}^H)$ as

$$\pi^{\mathsf{TTT}}_\theta(\cdot \mid x, y_{1:h-1}) := \pi_{\vartheta^{\mathsf{TTT}}(x, y_{1:h-1}; \theta)}(\cdot \mid x, y_{1:h-1}). \tag{64}$$

The distribution $\pi^{\mathsf{TTT}}_\theta$ can be interpreted as an augmented version of the autoregressive linear model π_θ that performs test-time training during sampling.

Proof. We closely follow the proof of Proposition 5.1 (cf. Appendix K.1).

We first note that by the proof of Eq. (57), we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\left\langle -\nabla \log \pi_{\theta^{t,h}}(y_h^t \mid x^t, y_{1:h-1}^t), \theta^{t,h} - \theta^\star \right\rangle\right] \le \frac{1}{2\eta} + \eta\,\mathbb{E}\left[\sum_{t=1}^{T}\left\|\nabla \log \pi_{\theta^{t,h}}(y_h^t \mid x^t, y_{1:h-1}^t)\right\|^2\right].$$

Further, by the proof of Eq. (58), we have

$$\begin{aligned}
&\left\|\nabla \log \pi_{\theta^{t,h}}(y_h^t \mid x^t, y_{1:h-1}^t)\right\|^2 \\
&\le 2\left\|\nabla \log \pi_{\theta^\star}(y_h^t \mid x^t, y_{1:h-1}^t)\right\|^2 \\
&\quad + 2B^2\left\langle \nabla \log \pi_{\theta^\star}(y_h^t \mid x^t, y_{1:h-1}^t) - \nabla \log \pi_{\theta^{t,h}}(y_h^t \mid x^t, y_{1:h-1}^t), \theta^{t,h} - \theta^\star \right\rangle
\end{aligned}$$

Note that the conditional distribution of $y_h^t \mid (x^t, y_{1:h-1}^t, \theta^{t,h})$ is given by $y_h^t \sim \pi_{\mathsf{D}}(\cdot \mid x^t, y_{1:h-1}^t)$. Hence, taking expectation, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla \log \pi_{\theta^{t,h}}(y_h^t \mid x^t, y_{1:h-1}^t)\right\|^2\right] &\le 2\,\mathbb{E}_{\pi_{\mathsf{D}}}\|\nabla \log \pi_{\theta^\star}(y_h \mid x, y_{1:h-1})\|^2 \\
&\quad + 2B^2\,\mathbb{E}\left[\left\langle -\nabla \log \pi_{\theta^{t,h}}(y_h^t \mid x^t, y_{1:h-1}^t), \theta^{t,h} - \theta^\star \right\rangle\right],
\end{aligned}$$

and we also have (cf. Eq. (55))

$$\mathbb{E}\left[\left\langle -\nabla \log \pi_{\theta^{t,h}}(y_h^t \mid x^t, y_{1:h-1}^t), \theta^{t,h} - \theta^\star \right\rangle\right] \ge \mathbb{E}\, D_{\mathsf{KL}}\left(\pi_{\mathsf{D}}(\cdot \mid x^t, y_{1:h-1}^t) \,\|\, \pi_{\theta^{t,h}}(\cdot \mid x^t, y_{1:h-1}^t)\right).$$

Combining the inequalities above, as long as $\eta \le \frac{1}{4B^2}$, it holds that

$$\mathbb{E}\left[\sum_{t=1}^{H}\sum_{h=1}^{H} D_{\mathsf{KL}}\left(\pi_{\mathsf{D}}(\cdot \mid x^t, y_{1:h-1}^t) \,\|\, \pi_{\theta^{t,h}}(\cdot \mid x^t, y_{1:h-1}^t)\right)\right] \le \frac{1}{\eta} + 4\eta T\sigma_\star^2. \tag{65}$$

Finally, we note that

$$\theta^{t,h} = \vartheta^{\mathsf{TTT}}(x^t, y_{h-1}^t; \theta^t),$$

and $x^t, y_{h-1}^t \mid \theta^t \sim \pi_{\mathsf{D}}$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[D_{\mathsf{KL}}\left(\pi_{\mathsf{D}}(\cdot \mid x^t, y_{1:h-1}^t) \,\|\, \pi_{\theta^{t,h}}(\cdot \mid x^t, y_{1:h-1}^t)\right) \mid \theta^t\right] \\
&= \mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}} D_{\mathsf{KL}}\left(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}) \,\|\, \pi_{\vartheta^{\mathsf{TTT}}(x, y_{1:h-1}; \theta^t)}(\cdot \mid x, y_{1:h-1})\right) \\
&= \mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}} D_{\mathsf{KL}}\left(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}) \,\|\, \pi_{\theta^t}^{\mathsf{TTT}}(\cdot \mid x, y_{1:h-1})\right).
\end{aligned}$$

This implies

$$\begin{aligned}
\frac{1}{\eta} + 4\eta T\sigma_\star^2 &\ge \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=1}^{H} D_{\mathsf{KL}}\left(\pi_{\mathsf{D}}(\cdot \mid x^t, y_{1:h-1}^t) \,\|\, \pi_{\theta^{t,h}}(\cdot \mid x^t, y_{1:h-1}^t)\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\sum_{h=1}^{H} D_{\mathsf{KL}}\left(\pi_{\mathsf{D}}(\cdot \mid x^t, y_{1:h-1}^t) \,\|\, \pi_{\theta^{t,h}}(\cdot \mid x^t, y_{1:h-1}^t)\right) \,\middle|\, \theta^t\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\pi_{\mathsf{D}}}\left[\sum_{h=1}^{H} D_{\mathsf{KL}}\left(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}) \,\|\, \pi_{\theta^t}^{\mathsf{TTT}}(\cdot \mid x, y_{1:h-1})\right)\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} D_{\mathsf{KL}}\left(\pi_{\mathsf{D}} \,\|\, \pi_{\theta^t}^{\mathsf{TTT}}\right)\right],
\end{aligned}$$

where the last equality uses the chain rule of KL divergence. □

L.2 Proof of Theorem F.2

We first note that by the proof of Eq. (57), we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\left\langle \mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}[\widehat{g}_{\theta^t}(y \mid x)], \theta^t - \theta^\star\right\rangle\right] \le \frac{1}{2\eta} + \eta\,\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}\|\widehat{g}_{\theta^t}(y \mid x)\|^2\right].$$

In the following, we analyze $\left\langle \mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}[\widehat{g}_{\theta}(y \mid x)], \theta^t - \theta^\star\right\rangle$ and $\mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}\|\widehat{g}_{\theta^t}(y \mid x)\|^2$ for any $\theta \in \Theta$, following the proof of Proposition 5.1 (cf. Appendix K.1).

We adopt the notation of Appendix K: For any θ and any pair $(x, y_{1:h-1})$, we denote $\bar{\phi}_\theta(x, y_{1:h-1}) = \mathbb{E}_{\pi_\theta}[\phi(x, y_{1:h}) \mid x, y_{1:h-1}]$ and

$$\epsilon_\theta(x, y_{1:h-1}) = D_{\mathsf{KL}}(\pi_{\mathsf{D}}(\cdot \mid x, y_{1:h-1}) \,\|\, \pi_\theta(\cdot \mid x, y_{1:h-1})).$$

By definition, we have (cf. Lemma K.4)

$$\sum_{h=1}^{H}\alpha_\theta(x, y_{1:h-1})\epsilon_\theta(x, y_{1:h-1}) = \min\left\{A, \sum_{h=1}^{H}\epsilon_\theta(x, y_{1:h-1})\right\}, \tag{66}$$

and hence

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}\left[\sum_{h=1}^{H}\alpha_\theta(x, y_{1:h-1})\epsilon_\theta(x, y_{1:h-1})\right] &= \mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}\min\left\{A, \sum_{h=1}^{H}\epsilon_\theta(x, y_{1:h-1})\right\} \\ &= D_{\mathsf{seq},N}(\pi_{\mathsf{D}} \,\|\, \pi_\theta),\end{aligned} \tag{67}$$

where we recall that $D_{\mathsf{seq},N}(\pi_{\mathsf{D}} \,\|\, \pi_\theta)$ is defined in Proposition D.10 and we denote $A = \log N$. Hence, by convexity (Eq. (55)),

$$\begin{aligned}&\left\langle \mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}[\widehat{g}_{\theta}(y \mid x)], \theta - \theta^\star\right\rangle \\ &= \mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}\left[\sum_{h=1}^{H}\alpha_\theta(x, y_{1:h-1})\left\langle \bar{\phi}_{\theta^\star}(x, y_{1:h-1}) - \bar{\phi}_\theta(x, y_{1:h-1}), \theta - \theta^\star\right\rangle\right] \\ &\ge \mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}\left[\sum_{h=1}^{H}\alpha_\theta(x, y_{1:h-1})\epsilon_\theta(x, y_{1:h-1})\right] = D_{\mathsf{seq},N}(\pi_{\mathsf{D}} \,\|\, \pi_\theta).\end{aligned}$$

Further, by Eq. (56), it holds that

$$\begin{aligned}&\|\widehat{g}_\theta(y \mid x) - \widehat{g}_{\theta^\star}(y \mid x)\| \\ &\le \sum_{h=1}^{H}\alpha_\theta(x, y_{1:h-1})\left\|\bar{\phi}_{\theta^\star}(x, y_{1:h}) - \bar{\phi}_\theta(x, y_{1:h-1})\right\| \\ &\le \sum_{h=1}^{H}\alpha_\theta(x, y_{1:h-1})\left(2\sqrt{\mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\cdot\epsilon_\theta(x, y_{1:h-1})} + 3B\epsilon_\theta(x, y_{1:h-1})\right).\end{aligned}$$

Hence, using Eq. (66), we have

$$\begin{aligned}&\|\widehat{g}_\theta(y \mid x) - \widehat{g}_{\theta^\star}(y \mid x)\|^2 \\ &\le 8\left(\sum_{h=1}^{H}\alpha_\theta(x, y_{1:h-1})\sqrt{\mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\cdot\epsilon_\theta(x, y_{1:h-1})}\right)^2 \\ &\quad + 18B^2\left(\sum_{h=1}^{H}\alpha_\theta(x, y_{1:h-1})\epsilon_\theta(x, y_{1:h-1})\right)^2 \\ &\le 8\left(\sum_{h=1}^{H}\alpha_\theta(x, y_{1:h-1})\mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\right)\left(\sum_{h=1}^{H}\alpha_\theta(x, y_{1:h-1})\epsilon_\theta(x, y_{1:h-1})\right) \\ &\quad + 18AB^2\left(\sum_{h=1}^{H}\alpha_\theta(x, y_{1:h-1})\epsilon_\theta(x, y_{1:h-1})\right) \\ &\le 8A\left(\sum_{h=1}^{H}\mathrm{Var}_{\pi_{\mathsf{D}}}(x, y_{1:h-1})\right) + 18AB^2\left(\sum_{h=1}^{H}\alpha_\theta(x, y_{1:h-1})\epsilon_\theta(x, y_{1:h-1})\right).\end{aligned}$$

Therefore, taking expectation of $(x, y) \sim \pi_{\mathsf{D}}$ and using $\mathbb{E}_{\pi_{\mathsf{D}}}\|\widehat{g}_{\theta^\star}(y \mid x)\|^2 \le \sigma_\star^2$ and Eq. (67), it holds that

$$\mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}\|\widehat{g}_\theta(y \mid x)\|^2 \le (16A+1)\sigma_\star^2 + 36AB^2 D_{\mathsf{seq},N}(\pi_{\mathsf{D}} \,\|\, \pi_\theta), \quad \forall\theta.$$

Finally, combining the inequalities above, we know that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T D_{\mathsf{seq},N}(\pi_{\mathsf{D}} \,\|\, \pi_{\theta^t})\right] &\le \mathbb{E}\left[\sum_{t=1}^T \left\langle \mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}[\widehat{g}_{\theta^t}(y \mid x)], \theta^t - \theta^\star \right\rangle\right] \\
&\le \frac{1}{2\eta} + \eta\,\mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_{(x,y)\sim\pi_{\mathsf{D}}}\|\widehat{g}_{\theta^t}(y \mid x)\|^2\right] \\
&\le \frac{1}{2\eta} + \eta T(16A+1)\sigma_\star^2 + 36AB^2\,\mathbb{E}\left[\sum_{t=1}^T D_{\mathsf{seq},N}(\pi_{\mathsf{D}} \,\|\, \pi_{\theta^t})\right].
\end{aligned}$$

Therefore, as long as $\eta \le \frac{1}{72AB^2}$, it holds that

$$\mathbb{E}\left[\sum_{t=1}^T D_{\mathsf{seq},N}(\pi_{\mathsf{D}} \,\|\, \pi_{\theta^t})\right] \le \frac{1}{\eta} + 36\eta T A\sigma_\star^2.$$

Optimally choosing η gives

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T D_{\mathsf{seq},N}(\pi_{\mathsf{D}} \,\|\, \pi_{\theta^t})\right] \lesssim \sqrt{\frac{\sigma_\star^2 \log N}{T}} + \frac{B^2 \log N}{T}.$$

By Proposition D.10, this implies

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \mathsf{Cov}_N(\pi_{\theta^t})\right] \lesssim \sqrt{\frac{\sigma_\star^2}{T\log N}} + \frac{B^2}{T}.$$

L.3 Proofs from Section 6.2 (Selection)

Below we state and prove a generalization of Theorem 6.2 which holds when the data distribution π_{D} is not necessarily in the model class Π.

Theorem 6.2′ (General version of Theorem 6.2)**.** *Fix $N \ge 1$, and consider the estimator $\widehat{\pi}$ from Eq. (13):*

$$\widehat{\pi} := \arg\min_{\pi\in\Pi} \max_{\pi'\in\Pi} \widehat{\mathsf{Cov}}_N(\pi' \,\|\, \pi). \tag{68}$$

For any $\delta \in (0,1)$, parameter $a, c \ge 0$, with probability at least $1-\delta$, it holds that

$$\mathsf{Cov}_{N^{1+a+2c}}(\widehat{\pi}) \lesssim \min_{\overline{\pi}\in\Pi} \mathsf{Cov}_{N^a}(\overline{\pi}) + \frac{1}{N^{1-a-2c}} + \frac{\log \mathcal{N}_\infty(\Pi; c\log N) + \log\delta^{-1}}{n}. \tag{69}$$

Proof of Theorem 6.2′. Fix any $\overline{\pi} \in \Pi$ and $M, \alpha > 0$, and we study the estimator

$$\widehat{\pi} := \arg\min_{\pi\in\Pi} \max_{\pi'\in\Pi} \widehat{\mathsf{Cov}}_M(\pi' \,\|\, \pi). \tag{70}$$

By Lemma J.2, with probability at least $1-\delta$, it holds that for $\forall \pi \in \Pi$,

$$\widehat{\mathsf{Cov}}_N(\overline{\pi} \,\|\, \pi) \ge \frac{1}{2}\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{2\alpha}N}(\overline{\pi} \,\|\, \pi) - \varepsilon_{\mathsf{stat}},$$

where $\varepsilon_{\mathsf{stat}} = \log(\mathcal{N}_\infty(\Pi;\alpha)/\delta)$. Next, again by Lemma J.2, with probability at least $1-\delta$, it holds that for $\forall \pi \in \Pi$,

$$\widehat{\mathsf{Cov}}_N(\pi \,\|\, \overline{\pi}) \le 2\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{-2\alpha}N}(\pi \,\|\, \overline{\pi}) + \varepsilon_{\mathsf{stat}}.$$

Therefore, we have with probability at least $1-2\delta$,

$$\begin{aligned}
\frac{1}{2}\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{2\alpha}N}(\overline{\pi} \,\|\, \widehat{\pi}) - \varepsilon_{\mathsf{stat}} &\le \widehat{\mathsf{Cov}}_N(\overline{\pi} \,\|\, \widehat{\pi}) \le \max_{\pi'\in\Pi} \widehat{\mathsf{Cov}}_N(\pi' \,\|\, \widehat{\pi}) \\
&= \min_{\pi\in\Pi}\max_{\pi'\in\Pi} \widehat{\mathsf{Cov}}_N(\pi' \,\|\, \pi) \le \max_{\pi'\in\Pi} \widehat{\mathsf{Cov}}_N(\pi' \,\|\, \overline{\pi}) \\
&\le 2\max_{\pi'\in\Pi} \mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{-2\alpha}N}(\pi \,\|\, \overline{\pi}) + \varepsilon_{\mathsf{stat}}.
\end{aligned}$$

Reorganizing yields

$$\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{2\alpha}N}(\overline{\pi} \,\|\, \widehat{\pi}) \le 4 \max_{\pi \in \Pi} \mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{-2\alpha}N}(\pi \,\|\, \overline{\pi}) + 4\varepsilon_{\mathsf{stat}}.$$

Note that for any N', N'' and policy π, π', π'',

$$\mathsf{Cov}^{\pi_{\mathsf{D}}}_{N'N''}(\pi' \,\|\, \pi) \le \mathsf{Cov}^{\pi_{\mathsf{D}}}_{N'}(\pi' \,\|\, \pi'') + \mathsf{Cov}^{\pi_{\mathsf{D}}}_{N''}(\pi'' \,\|\, \pi). \tag{71}$$

Hence, for all π,

$$\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{2\alpha}NN'}(\pi_{\mathsf{D}} \,\|\, \pi) \le \mathsf{Cov}^{\pi_{\mathsf{D}}}_{N'}(\pi_{\mathsf{D}} \,\|\, \overline{\pi}) + \mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{2\alpha}N}(\overline{\pi} \,\|\, pi),$$
$$\mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{-2\alpha}N}(\pi \,\|\, \overline{\pi}) \le \mathsf{Cov}^{\pi_{\mathsf{D}}}_{N'}(\pi \,\|\, \pi_{\mathsf{D}}) + \mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{-2\alpha}N/N'}(\pi_{\mathsf{D}} \,\|\, \overline{\pi})$$

Therefore, using the fact that $\mathsf{Cov}^{\pi_{\mathsf{D}}}_{A}(\pi \,\|\, \pi_{\mathsf{D}}) \le \frac{1}{A}$ and the inequalities above, we see that

$$\begin{aligned}
\mathsf{Cov}_{e^{2\alpha}NN'}(\widehat{\pi}) &= \mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{2\alpha}NN'}(\pi_{\mathsf{D}} \,\|\, \widehat{\pi}) \\
&\le \mathsf{Cov}^{\pi_{\mathsf{D}}}_{N'}(\pi_{\mathsf{D}} \,\|\, \overline{\pi}) + \mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{2\alpha}N}(\overline{\pi} \,\|\, \widehat{\pi}) \\
&\le \mathsf{Cov}^{\pi_{\mathsf{D}}}_{N'}(\pi_{\mathsf{D}} \,\|\, \overline{\pi}) + 4 \max_{\pi \in \Pi} \mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{-2\alpha}N}(\pi \,\|\, \overline{\pi}) + 4\varepsilon_{\mathsf{stat}} \\
&\le 5\mathsf{Cov}^{\pi_{\mathsf{D}}}_{N'}(\pi_{\mathsf{D}} \,\|\, \overline{\pi}) + 4 \max_{\pi \in \Pi} \mathsf{Cov}^{\pi_{\mathsf{D}}}_{e^{-2\alpha}N/N'}(\pi \| \pi_{\mathsf{D}}) + 4\varepsilon_{\mathsf{stat}} \\
&\le 5\mathsf{Cov}^{\pi_{\mathsf{D}}}_{N'}(\pi_{\mathsf{D}} \,\|\, \overline{\pi}) + \frac{e^{2\alpha}N'}{N} + 4\varepsilon_{\mathsf{stat}}.
\end{aligned}$$

The claimed bound follows by setting $\overline{\pi} = \arg\min_{\pi \in \Pi} \mathsf{Cov}^{\pi_{\mathsf{D}}}_{N'}(\pi_{\mathsf{D}} \,\|\, \overline{\pi})$, $\alpha = c \log N$, and $N' = N^a$. □

L.4 Proof of Theorem F.3

Divergence. For any distribution $P, Q \in \Delta(\mathcal{Y})$, we define the following divergence for $M \ge 1$:

$$\mathcal{E}_M(P \,\|\, Q) := \max\left\{ \mathbb{E}_{y \sim P}\left(\frac{dQ}{dP} - M\right)_+, \mathbb{E}_{y \sim Q}\left(\frac{dP}{dQ} - M\right)_+ \right\}.$$

Then, for policies $\pi, \pi' : \mathcal{X} \to \Delta(\mathcal{Y})$, we further define

$$\mathcal{E}_{M,\mu}(\pi \,\|\, \pi') := \mathbb{E}_{x \sim \mu} \mathcal{E}_M(\pi(\cdot \mid x) \,\|\, \pi'(\cdot \mid x)).$$

Under this divergence, it holds that for any event E,

$$\mathbb{P}_{\mu,\pi}(E) \le M \cdot \mathbb{P}_{\mu,\pi'}(E) + \mathcal{E}_{M,\mu}(\pi \,\|\, \pi'), \tag{72}$$
$$\mathbb{P}_{\mu,\pi'}(E) \le M \cdot \mathbb{P}_{\mu,\pi}(E) + \mathcal{E}_{M,\mu}(\pi \,\|\, \pi'), \tag{73}$$

where $\mathbb{P}_{\mu,\pi}$ is the probability under $x \sim \mu$ and $y \sim \pi(\cdot \mid x)$. Furthermore, we can bound

$$\mathsf{Cov}_{2M}(\pi) = \mathbb{P}_{\mu,\pi_{\mathsf{D}}}\left(\frac{\pi_{\mathsf{D}}(y \mid x)}{\pi(y \mid x)} \ge 2M\right) \le \mathcal{E}_{M,\mu}(\pi_{\mathsf{D}} \,\|\, \pi). \tag{74}$$

Theorem F.3′ (General version of Theorem F.3)**.** *Fix $N, \gamma \ge 1$ such that $N \ge 4\gamma^2$. Consider the estimator*

$$\widehat{\pi} := \arg\min_{\pi \in \Pi} \max_{\pi' \in \Pi} \left\{ \widehat{\mathsf{Cov}}_N(\pi' \,\|\, \pi) - 2\gamma \cdot \widehat{\mathsf{Cov}}^{\pi}_N(\pi' \,\|\, \pi) \right\}. \tag{75}$$

Then with probability $1 - \delta$, it holds that

$$\mathsf{Cov}_{2N\gamma}(\widehat{\pi}) \lesssim \min_{\pi \in \Pi} \mathcal{E}_\gamma(\pi_{\mathsf{D}} \,\|\, \pi) + \frac{\log(|\Pi|/\delta)}{n}.$$

Theorem F.3 is an immediate corollary by setting $\gamma = N^c$.

Proof of Theorem F.3′. For $\pi, \pi' \in \Pi$, we define the set

$$\mathcal{C}_N(\pi, \pi') = \left\{ (x, y) \mid \frac{\pi(y \mid x)}{\pi'(y \mid x)} \ge N \right\}.$$

Suppose an i.i.d. dataset $\mathcal{D} = \{(x^i, y^i)\}_{i\in[n]} \sim \pi_{\mathsf{D}}$ is drawn. We write $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{(x^i,y^i)}$ to denote the empirical measure (i.e., μ_n is the uniform distribution over $\mathcal{D}$). Note that

$$\widehat{\mathsf{Cov}}_N(\pi' \,\|\, \pi) = \mu_n(\mathcal{C}_N(\pi', \pi)). \tag{76}$$

We also recall that

$$\widehat{\mathsf{Cov}}^{\overline{\pi}}_N(\pi' \,\|\, \pi) := \frac{1}{n}\sum_{i=1}^n \mathbb{P}_{y\sim\overline{\pi}(\cdot|x^i)}\left(\frac{\pi'(y \mid x^i)}{\pi(y \mid x^i)} \geq N\right).$$

Therefore, we write $\widehat{\mathsf{Cov}}^{\overline{\pi}}_N(\pi' \,\|\, \pi) = \mathbb{P}_{n,\overline{\pi}}(\mathcal{C}_N(\pi', \pi))$, where $\mathbb{P}_{n,\overline{\pi}}$ is the probability under the distribution $x \sim \mu_n, y \sim \overline{\pi}(\cdot \mid x)$.

Thus, the tournament estimator in Eq. (38) can be expressed as

$$\widehat{\pi} := \arg\min_{\pi\in\Pi} \max_{\pi'\in\Pi} \mathcal{L}(\pi, \pi'), \tag{77}$$

where

$$\mathcal{L}(\pi, \pi') := \mu_n(\mathcal{C}_N(\pi', \pi)) - 2\gamma \cdot \mathbb{P}_{n,\overline{\pi}}(\mathcal{C}_N(\pi', \pi)), \tag{78}$$

and $\gamma = 2N^c$.

As an immediate consequence of Lemma H.3 and union bound, we have the following:

Lemma L.1. *Fix $\delta \in (0,1)$, and define $\varepsilon_{\mathsf{stat}} = \frac{16\log(4|\Pi|/\delta)}{n}$. With probability $1-\delta$, the following holds simultaneously:*

(1) For all $\pi, \pi' \in \Pi$, it holds that

$$2\mathbb{P}_{\mu,\pi_{\mathsf{D}}}(\mathcal{C}_N(\pi', \pi)) + \varepsilon_{\mathsf{stat}} \geq \mu_n(\mathcal{C}_N(\pi', \pi)) \geq \frac{1}{2}\mathbb{P}_{\mu,\pi_{\mathsf{D}}}(\mathcal{C}_N(\pi', \pi)) - \varepsilon_{\mathsf{stat}},$$
$$2\mathbb{P}_{n,\pi_{\mathsf{D}}}(\mathcal{C}_N(\pi', \pi)) + \varepsilon_{\mathsf{stat}} \geq \mu_n(\mathcal{C}_N(\pi', \pi)) \geq \frac{1}{2}\mathbb{P}_{n,\pi_{\mathsf{D}}}(\mathcal{C}_N(\pi', \pi)) - \varepsilon_{\mathsf{stat}}.$$

(2) For any $\pi \in \Pi$, it holds that $\mathcal{E}_{\gamma,\mu_n}(\pi_{\mathsf{D}} \,\|\, \pi) \leq 2\mathcal{E}_{\gamma,\mu}(\pi_{\mathsf{D}} \,\|\, \pi) + \varepsilon_{\mathsf{stat}}$.

In the following, we fix $\delta \in (0,1)$ and condition on the success event of Lemma L.1. Let $\overline{\pi} \in \Pi$ denote some policy for which $\varepsilon_{\mathsf{apx}} = \mathcal{E}_{\gamma,\mu}(\pi_{\mathsf{D}} \,\|\, \overline{\pi})$. We denote $\varepsilon'_{\mathsf{apx}} = \mathcal{E}_{\gamma,\mu_n}(\pi_{\mathsf{D}} \,\|\, \overline{\pi})$, and by Lemma L.1, we have $\varepsilon'_{\mathsf{apx}} \leq 2\varepsilon_{\mathsf{apx}} + \varepsilon_{\mathsf{stat}}$.

Then, for any $\pi' \in \Pi$,

$$\begin{aligned}\mathcal{L}(\overline{\pi}, \pi') &\leq 2\mathbb{P}_{n,\pi_{\mathsf{D}}}(\mathcal{C}_N(\pi', \overline{\pi})) - 2\gamma\mathbb{P}_{n,\overline{\pi}}(\mathcal{C}_N(\pi', \overline{\pi})) + \varepsilon_{\mathsf{stat}} \\ &\leq 2\mathcal{E}_{\gamma,\mu_n}(\pi_{\mathsf{D}} \,\|\, \overline{\pi}) + \varepsilon_{\mathsf{stat}} = \varepsilon'_{\mathsf{apx}} + \varepsilon_{\mathsf{stat}}.\end{aligned}$$

where the first inequality uses Lemma L.1, and the second inequality uses Eq. (72).

Therefore, we have

$$\max_{\pi'\in\Pi} \mathcal{L}(\widehat{\pi}, \pi') = \min_{\pi\in\Pi}\max_{\pi'\in\Pi} \mathcal{L}(\pi, \pi') \leq \max_{\pi'\in\Pi} \mathcal{L}(\overline{\pi}, \pi') \leq \varepsilon_{\mathsf{stat}} + \varepsilon'_{\mathsf{apx}}.$$

In particular, we know $\mathcal{L}(\widehat{\pi}, \overline{\pi}) \leq \varepsilon_{\mathsf{stat}} + \varepsilon'_{\mathsf{apx}}$. Then, we can bound

$$\begin{aligned}\mu_n(\mathcal{C}_N(\overline{\pi}, \widehat{\pi})) - \mathcal{L}(\widehat{\pi}, \overline{\pi}) &= 2\gamma\mathbb{P}_{n,\widehat{\pi}}(\mathcal{C}_N(\overline{\pi}, \widehat{\pi})) \\ &\leq \frac{2\gamma}{N}\mathbb{P}_{n,\overline{\pi}}(\mathcal{C}_N(\overline{\pi}, \widehat{\pi})) \\ &\leq \frac{2\gamma}{N}\left[\frac{\gamma}{2}\mathbb{P}_{n,\pi_{\mathsf{D}}}(\mathcal{C}_N(\overline{\pi}, \widehat{\pi})) + \varepsilon'_{\mathsf{apx}}\right] \\ &\leq \frac{2\gamma}{N}\left[\gamma(\mu_n(\mathcal{C}_N(\overline{\pi}, \widehat{\pi})) + \varepsilon_{\mathsf{stat}}) + \varepsilon'_{\mathsf{apx}}\right],\end{aligned}$$

where the second inequality uses Eq. (73): $\mathbb{P}_{n,\overline{\pi}}(E) - \frac{\gamma}{2}\mathbb{P}_{n,\pi_{\mathsf{D}}}(E) \leq \mathcal{E}_{M,\mu_n}(\pi_{\mathsf{D}} \,\|\, \overline{\pi}) = \varepsilon'_{\mathsf{apx}}$ for any event E, and the thrid inequality uses Lemma L.1. Therefore, using $N \geq 4\gamma^2$, we know $\mu_n(\mathcal{C}_N(\overline{\pi}, \widehat{\pi})) \leq 5\varepsilon_{\mathsf{stat}} + 2\varepsilon'_{\mathsf{apx}}$. Using Lemma L.1 again, we have

$$\mathsf{Cov}^{\pi_{\mathsf{D}}}_N(\overline{\pi} \,\|\, \widehat{\pi}) = \mathbb{P}_{\mu,\pi_{\mathsf{D}}}(\mathcal{C}_N(\overline{\pi}, \widehat{\pi})) \leq 2\mu_n(\mathcal{C}_N(\overline{\pi}, \widehat{\pi})) + 2\varepsilon_{\mathsf{stat}} \leq 12\varepsilon_{\mathsf{stat}} + 4\varepsilon'_{\mathsf{apx}}.$$

By Eq. (71), it holds that

$$\mathsf{Cov}_{2N\gamma}(\pi) = \mathsf{Cov}^{\pi_{\mathsf{D}}}_{2N\gamma}(\pi_{\mathsf{D}} \,\|\, \pi) \le \mathsf{Cov}^{\pi_{\mathsf{D}}}_{2\gamma}(\pi_{\mathsf{D}} \,\|\, \overline{\pi}) + \mathsf{Cov}^{\pi_{\mathsf{D}}}_{N}(\overline{\pi} \,\|\, \pi),$$

and we also have $\mathsf{Cov}^{\pi_{\mathsf{D}}}_{2\gamma}(\pi_{\mathsf{D}} \,\|\, \overline{\pi}) \le \varepsilon_{\mathsf{apx}}$ by Eq. (74). Combining the inequalities above, we can conclude that

$$\mathsf{Cov}_{2N\gamma}(\widehat{\pi}) \le \mathsf{Cov}^{\pi_{\mathsf{D}}}_{N}(\overline{\pi} \,\|\, \pi) + \varepsilon_{\mathsf{apx}} \le 12\varepsilon_{\mathsf{stat}} + 4\varepsilon'_{\mathsf{apx}} + \varepsilon_{\mathsf{apx}}.$$

Finally, using Lemma H.3, we have $\varepsilon'_{\mathsf{apx}} \le 2\varepsilon_{\mathsf{apx}} + \varepsilon_{\mathsf{stat}}$. This is the desired upper bound. □