REINFORCEMENT LEARNING FOR HIERARCHICAL REASONING REWARDS: MINIMAX-OPTIMAL RATES WITH TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has become a standard tool for post-training language models on reasoning tasks, where the policy is updated by reward feedback while exploring the space of responses. Despite its empirical success, theoretical understanding of RL post-training remains limited, in particular of why on-policy exploration combined with a neural reward model is effective. In this paper, we address this question by modeling the reward as a hierarchical function on the response space: the reward consists of infinitely many local components, each of which becomes relevant only after the preceding ones have been resolved. We show that a natural Transformer-based actor–critic algorithm, which alternates between sampling from the current KL-regularized policy, fitting a Transformer critic to the observed rewards, and updating the policy, achieves the minimax optimal rates in the query budget and in the regularization strength up to logarithmic factors, and is minimax optimal for a fixed number of prompts. In contrast, we prove that sampling from the fixed reference distribution, as in offline reward modeling, can limit regret decay to a logarithmic rate. These results show that on-policy exploration progressively zooms in on the region where the reward is concentrated, and quantify its benefit for RL post-training.

1 INTRODUCTION

Reinforcement learning (RL) has emerged as a central ingredient in post-training large language models (LLMs) for complex reasoning. Recent systems have demonstrated substantial gains on mathematical reasoning, coding, and other verifiable tasks by optimizing model-generated responses against outcome-based rewards (Shao et al., 2024; Guo et al., 2025; Kimi Team, 2025). In contrast to supervised fine-tuning from a fixed collection of demonstrations, RL post-training repeatedly generates responses from the current policy, scores them with a reward model or verifier, and updates the policy with the resulting feedback. This creates a feedback loop: as the model improves, the distribution of responses from which subsequent training data are collected changes as well.

Despite the empirical success of this paradigm, the statistical role of such adaptive data collection is not yet fully understood. Recent theoretical work has identified several benefits of exploration in LLM reasoning and alignment, through the dynamics of reasoning, the generalization of the learned policy, or the implicit exploration of a KL-regularized policy (Foster et al., 2025; Kim et al., 2025; Wang et al., 2026a;b), and suggests that *where* the training responses come from can be as important as the learning rule applied to them. Yet a basic statistical question remains open: *when and why does adapting the response distribution improve the rate of learning*, in particular for the reasoning tasks that have driven recent progress in RL post-training, in which a correct solution must clear several difficult steps in succession?

To study this question, we formulate LLM post-training as KL-regularized reward maximization over a response space and introduce a hierarchical model of reasoning rewards. The motivation is that complex reasoning rarely reveals all of its difficulties at once: only after one hard step has been cleared does it become apparent that another, again nontrivial, problem must be solved next. In a mathematical proof, for example, one may first need to find the right theorem to invoke; only once this is settled does it become clear which transformation is needed to apply it, and only then do the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

finer algebraic and logical decisions come into view. We model this by a hierarchy of smooth reward components supported on nested regions of the response space, with progressively more localized reward structure appearing closer to optimal responses.

This hierarchical structure creates a sharp statistical distinction between fixed and adaptive response collection. Under the uniform reference distribution, the nested regions carrying deeper reward components occupy exponentially small fractions of the response space, and observations outside a region carry no information about the component inside it: one cannot tell that a deeper step is needed before getting that far. Learning a deep component is therefore a rare-event problem. On-policy RL removes this barrier: once a coarse reward component has been learned, the policy shifts its mass toward the corresponding promising region, so that the next batch is collected precisely where the next, finer component becomes visible. Repeating this step progressively localizes the data distribution around high-reward responses.

Contributions. We make this intuition precise through a minimax analysis. Let n denote the total number of reward observations, M the number of prompts, $\rho \geq 0$ the strength of the KL regularization, $\gamma > 1$ the rate at which deeper reward components decay, and α the smoothness of the components in a d-dimensional response embedding space. Our contributions are threefold.

1. **Hierarchical reward model and minimax characterization.** We introduce the nested hierarchical reward model and characterize the minimax regret over all (adaptive) algorithms. For fixed M, when the regularization is absent or weak, the minimax rate is $\widetilde{\Theta}\big(n^{-(\gamma-1)/(2\gamma+1)}\big)$: the difficulty is how deep in the hierarchy the learner can explore. In the large-budget regime, the rate is $\widetilde{\Theta}\big(\rho\,(n/n_0)^{-2\alpha/(2\alpha+d)}\big)$ with $n_0 := M\rho^{-(2+1/\gamma)}$: the difficulty is the local nonparametric estimation of the reward near the optimum. The upper bounds match the lower bounds up to logarithmic factors and a factor polynomial in M, and up to logarithmic factors for fixed M.

2. **A minimax-optimal Transformer actor–critic.** We give an on-policy actor–critic whose critic is a causal Transformer, refitted at every iteration to fresh on-policy observations, and whose actor is the resulting Gibbs policy. Such a procedure, motivated by standard RL post-training, attains both rates above up to logarithmic factors and is thus minimax optimal for fixed M.

3. **A statistical separation from fixed response collection.** We establish lower bounds for algorithms whose responses are drawn from the fixed reference distribution μ, as in offline reward modeling, while the estimator and the output policy are unrestricted. Because deeper reward components are exponentially rare under fixed sampling, the regret can decay only logarithmically in n/M when the regularization is weak. The benefit of on-policy RL is therefore statistical, not merely an optimization effect of the policy update.

Related work. A first line of work adapts the distribution of tasks or prompts during training: easy-to-hard schedules (Li et al., 2023; Parashar et al., 2026), adaptive prompt selection from the current performance (Rajaraman et al., 2026), and depth- or hint-based curricula (Bu et al., 2026). The last makes RL post-training with outcome-only rewards polynomial instead of exponential on a discrete reasoning tree. There the adaptivity is over prompts or tasks and the intermediate signal is supplied by the curriculum; here the prompt distribution is fixed and the learner adapts only its response distribution, which is possible because the reward is graded rather than outcome-only.

Technically, the closest algorithms are those for continuum-armed bandits, which partition the action space hierarchically and refine it around the maximizer (Kleinberg et al., 2008; Bubeck et al., 2011; Munos, 2011; Valko et al., 2013). We depart from that line in three ways, all motivated by LLM reasoning: the reward is hierarchical, with a new difficulty appearing at every level; the target is a diverse policy, the Gibbs policy at temperature $\rho > 0$, rather than the maximizer alone; and the optimal rates are attained by fitting an LLM-compatible Transformer critic and updating a Gibbs actor, rather than by optimistic tree search. Nishikawa & Suzuki (2026) likewise alternate between updating a policy and refitting a neural reward estimator, but for a non-hierarchical reward with a fully connected estimator that must be smoothed before it drives the policy, and their rate is not matched by a lower bound. Appendix A discusses these two lines in more detail, together with search over reasoning trajectories, on-policy exploration under KL regularization, and neural function approximation in estimation and RL.

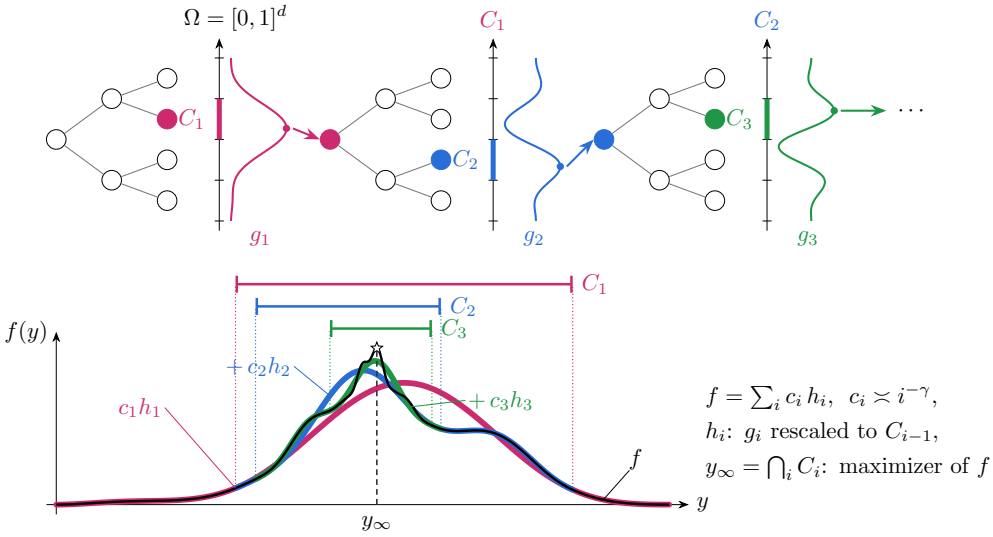


Figure 1: A hierarchically nested reward (Definition 1), for $d = 1$ and zoom depths $\ell_i = 2$. (Top) A response is a path in a tree with 2^d children per node, and a prefix of depth h is a dyadic cube of that depth. The primitive g_1 lives on Ω; inside the cell C_1 around its maximizer, g_2 is added, rescaled to C_1, and so on. (Bottom) The resulting reward $f = \sum_i c_i h_i$: bumps of decreasing weight $c_i \asymp i^{-\gamma}$ and decreasing width accumulate around the maximizer $y_\infty = \bigcap_i C_i$.

2 Problem Setting

We first introduce the class of hierarchically nested rewards. We then formulate RL post-training as KL-regularized maximization of such a reward. Finally, we describe how it can be queried.

2.1 Hierarchically Nested Rewards

We consider multi-step reasoning tasks with a hierarchical reward: at first only the coarse structure of the problem is visible, and solving one step reveals a further problem nested inside it. We model this structure by functions on the response embedding space $\Omega = [0,1]^d$ in which a new term is added each time the argument enters a small region around the maximizer of the previous term; how a response of a language model is mapped to a point of Ω is explained in Section 2.2.

Hierarchically nested functions. Let μ be the uniform distribution on Ω, $\mu(S)$ being the Lebesgue measure of $S \subset \mathbb{R}^d$ and $\mathbb{1}_S$ its indicator. A *dyadic cube* of depth $h \geq 0$ is a cube of side 2^{-h} of the form $\prod_{j \leq d} [k_j 2^{-h}, (k_j + 1)2^{-h})$ with integers $0 \leq k_j < 2^h$, and $D_h(y)$ is the depth-h dyadic cube containing y, so that $\Omega = D_0(y) \supset D_1(y) \supset \cdots$. The idea is that a first function g_1 is defined on the whole cube; once the argument enters the small dyadic cube around the maximizer of g_1, a second function g_2 is added on that cube (rescaled to Ω), and so on.

Definition 1 (Hierarchically nested function class)**.** Let $\mathcal{G} = (g_i)_{i \geq 1}$ be a sequence of uniformly bounded functions $g_i : \Omega \to \mathbb{R}$, let $c = (c_i)_{i \geq 1}$ be positive numbers with $\sum_i c_i < \infty$, and let $\ell = (\ell_i)_{i \geq 1}$ be positive integers (*zoom depths*); write $L_i = \sum_{a \leq i} \ell_a$, $L_0 = 0$. For each i, fix a maximizer $z_i^\star \in \arg\max_{u \in \Omega} g_i(u)$ (ties broken by a fixed rule) and let Q_i be the depth-ℓ_i dyadic cube containing $z_i^\star$. Define the *active cells* recursively by $C_0 = \Omega$ and $C_i = b_{i-1} + 2^{-L_{i-1}} Q_i$, of depth L_i, where b_{i-1} is the lower-left corner of C_{i-1}, so that $C_{i-1} = b_{i-1} + 2^{-L_{i-1}} \Omega$. The *hierarchically nested function* generated by $(\mathcal{G}, c, \ell)$ is

$$f(y) = \sum_{i=1}^{\infty} c_i \, h_i(y), \qquad h_i(y) = g_i^0\big(2^{L_{i-1}}(y - b_{i-1})\big), \tag{1}$$

where g_i^0 is the extension of g_i by zero to $\mathbb{R}^d$, so that h_i is a rescaled copy of g_i on C_{i-1} and vanishes outside. We write $\mathcal{H}(\mathcal{G}, c, \ell)$ for f, and $\mathcal{H}$ for the class of all such functions.

The cells are nested, $\Omega = C_0 \supset C_1 \supset \cdots$, and the i-th level $c_i h_i$ is visible only inside C_{i-1}: solving level i (entering C_i) is a prerequisite for collecting the rewards of the deeper levels. Figure 1 illustrates the construction.

We make the following assumption; see Appendix C.1 for α-Hölder smoothness.

Assumption 2. *Let $\alpha > 0$, $\lambda > 0$, $c_{\min} \in (0, 1)$, $\gamma > 1$ and integers $1 \le \ell_\star \le \bar{\ell} < \infty$ be known constants, and let $c_i = i^{-\gamma}/\zeta(\gamma)$ with $\zeta(\gamma) = \sum_{i \ge 1} i^{-\gamma}$. The function $f = \mathcal{H}(\mathcal{G}, c, \ell)$ has unknown primitives $\mathcal{G} = (g_i)_{i \ge 1}$ and unknown zoom depths $\ell = (\ell_i)_{i \ge 1}$ that satisfy the following conditions: **(A1)** (Smoothness and boundedness) $g_1 \in [0, 1]$, $g_i \in [-1, 1]$, and g_i^0 is α-Hölder smooth with constant λ for all i. **(A2)** (Positive maximum) $g_i(z_i^\star) \ge c_{\min}$ for all i. **(A3)** (Deep enough zoom) $\ell_i \in [\ell_\star, \bar{\ell}]$ for all i, and $3^\gamma 2^{-\beta \ell_\star} \le \min\{1/2, c_{\min}/(8\lambda)\}$. **(A4)** (Zero mean) $\int_\Omega g_i \, d\mu = 0$ ($i \ge 2$).*

We collect the *structural parameters* as $\theta = (d, \alpha, \lambda, c_{\min}, \gamma, \ell_\star, \bar{\ell})$, and write $T_k = \sum_{i > k} c_i \asymp k^{-(\gamma-1)}$ for the weight of the levels deeper than k. Constants such as C, c and those implicit in $a \lesssim b$, $a \asymp b$ and $\widetilde{O}(\cdot)$ depend only on θ and the noise level σ of Section 2.3.

Remark. **(A1)** is a standard smoothness condition. Requiring it of the zero extension g_i^0 makes a primitive vanish continuously at the boundary of its cell, so that f is continuous although it has infinitely many levels. **(A2)** and **(A3)** guarantee a genuine "correct continuation" at every level: C_i is so small that the variation of the previous levels inside it is dominated by the gain $c_{\min} c_i$ of solving level i, uniformly in i. **(A4)** makes the deeper levels cancel when averaged over a cell containing them, so that the average of f over a region (Section 2.3) reveals only the levels that this region has resolved; by analogy with mathematical research, *the deeper one digs into a problem, the deeper the results that come into view*. Under Assumption 2, $f \in [-1, 1]$ and its unique maximizer is the common point y_∞ of the nested cells (Lemma 6).

2.2 RL Post-Training as KL-Regularized Reward Maximization

Let $\mathcal{X}$ be the set of prompts and $\mathcal{Y}$ the set of responses. A *policy* P specifies, for every prompt x, a conditional distribution $P(\cdot \mid x)$ over responses; a language model is a policy, and generating a response means sampling from it. The quality of a response is measured by an unknown *reward function* $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Given a reference policy π_{ref}, typically the model before post-training, and a regularization strength $\rho \in [0, 1]$, post-training maximizes the KL-regularized objective

$$J_\rho(P; x) = \mathbb{E}_{y \sim P(\cdot|x)}[f(x, y)] - \rho \, \mathrm{KL}\left(P(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x)\right) \tag{2}$$

over prompts $x \sim \nu$, where $\mathrm{KL}(\cdot \,\|\, \cdot)$ is the Kullback–Leibler divergence; this is the objective of KL-regularized post-training such as PPO and GRPO (Ziegler et al., 2019; Ouyang et al., 2022; Schulman et al., 2017; Shao et al., 2024). The performance of a policy is its *regret*

$$J_\rho^*(x) := \sup_P J_\rho(P; x), \qquad \mathcal{R}_\rho(P) = \mathbb{E}_{x \sim \nu}\big[J_\rho^*(x) - J_\rho(P; x)\big]. \tag{3}$$

For $\rho = 0$ we have $J_0^*(x) = \sup_y f(x, y)$, so $\mathcal{R}_0$ is the unregularized regret. The learner does not know f and can only query a noisy reward oracle a limited number of times (Section 2.3), which is the dominant cost when the reward comes from a verifier or a human.

For simplicity, we take the prompts to be M labels, $\mathcal{X} = [M]$, drawn uniformly, $\nu = \mathrm{Unif}([M])$, and the responses to be infinite binary sequences, $\mathcal{Y} = \{0, 1\}^{\{0,1,2,\ldots\}}$, generated one bit at a time; the reference policy π_{ref} generates the bits uniformly at random. Prompts are exogenous: at every round of interaction, a fresh prompt is drawn from ν independently of the past, and the learner cannot choose or reject prompts. The reward functions of different prompts are unrelated, and we write $f(y; m)$ for $f(m, y)$.

Embedding of responses. We embed responses into $\Omega = [0, 1]^d$ so that the responses sharing a given prefix form a dyadic cube, and a longer prefix a smaller cube inside it. Concretely, we interleave the bits across the d coordinates in a round-robin fashion: $\phi : \mathcal{Y} \to \Omega$ with $\phi_j(y) = \sum_{k \ge 0} y_{kd+j-1} 2^{-(k+1)}$ for $j \in [d]$. A prefix of length dh then corresponds to the dyadic cube $D_h(\phi(y))$, and generating bits uniformly corresponds to $\phi(y) \sim \mu$. We identify a response with $\phi(y)$ and a prefix with a cube, so that a policy is a distribution over Ω. The embedding can be read as part of the language model itself: it places responses that share a long prefix close to each other,

so that the reward, as a function on Ω, records how the quality of a response depends on its reasoning prefix. Any embedding whose regions form a family of dyadic cubes would do; the results do not depend on the particular interleaving.

Hierarchically nested rewards of prompts. For every prompt $m \in [M]$ the reward is a hierarchically nested function $f(\cdot; m) = \mathcal{H}(\mathcal{G}_m, c, \ell)$ satisfying Assumption 2, with unknown primitives $\mathcal{G}_m = (g_{i,m})_{i \geq 1}$; the weights c and the zoom depths ℓ are shared, and we write $C_i(m)$, $h_i(\cdot; m)$, $y_\infty(m)$ for the cells, levels and maximizer of prompt m, and $\mathcal{F}_{M,\theta}$ for the resulting class of rewards. Each level is one difficulty of the task: its contribution $c_i \asymp i^{-\gamma}$ decays with the depth at which it appears and is collected only once the earlier ones have been resolved. The reward is thus a function of the whole response rather than of a final answer, as in post-training pipelines where a verifier or a generative reward model scores the entire trajectory (Shao et al., 2025; DeepSeek-AI, 2026).

2.3 Reward Oracle and Learning Protocol

The learner interacts with the environment for n rounds, each of which yields exactly one observation. At round t a prompt $X_t \sim \nu$ arrives, the learner generates a possibly partial response and queries the reward of *one* prefix, i.e. one dyadic cube D_t containing the response. The oracle returns the average reward over that cube plus noise,

$$Z_t = Q(X_t, D_t) + \xi_t, \qquad Q(m, D) := \mathbb{E}_{y \sim \mu}[f(y; m) \mid y \in D], \tag{4}$$

where, conditionally on the history and on (X_t, D_t), the noise is mean zero and σ-sub-Gaussian, with σ (or an upper bound on it) known to the learner.

Equivalently, $Q(m, D)$ is the expected terminal reward of a response with prefix D whose remaining bits are generated by the reference policy, the randomness of the completion being absorbed into the noise. This value can be estimated by Monte Carlo completions, as in methods that derive process supervision from terminal outcomes (Wang et al., 2024; Setlur et al., 2024). By **(A3)** and **(A4)**, a prefix of depth h resolves at most the first $\max\{1, \lceil h/\ell_\star \rceil\}$ levels, and deeper levels do not contribute to $Q(m, D)$.

After the n rounds the learner outputs a policy $\widehat{P}(\cdot \mid m)$ for every prompt. It knows θ and σ, but neither the primitives, nor the zoom depths, nor the cells, nor the maximizers.

3 Transformer Actor–Critic Attains the Minimax Rates

In this section, we present an actor–critic algorithm in which the critic is a Transformer trained by empirical risk minimization on the full class and the actor is a KL-regularized (Gibbs) policy, and prove its regret bound.

3.1 Transformer Critic

The critic takes a prompt and a prefix and outputs a scalar. Its input is a token sequence $X = (x_1, \ldots, x_T)$ over the vocabulary $\mathcal{V} = [M] \cup \{0, 1, \texttt{STOP}, \texttt{PAD}, \texttt{OUT}\}$: the prompt token m, the dh bits of the prefix, a `STOP` token, `PAD` tokens, and a readout token `OUT` at the last position, the `PAD` tokens bringing every input to the same length T. Tokens are represented by one-hot vectors, $X \in \mathbb{R}^{|\mathcal{V}| \times T}$. Following Takakura & Suzuki (2023), the Transformer architecture has three components: (i) a position-wise FNN layer, (ii) a self-attention layer, and (iii) an embedding layer.

(i) An FNN of depth L and width W is $f(x) := (A_L \eta(\cdot) + b_L) \circ \cdots \circ (A_1 x + b_1)$ with $A_i \in \mathbb{R}^{d_{i+1} \times d_i}$, $b_i \in \mathbb{R}^{d_{i+1}}$, $\max_i d_i \leq W$ and the ReLU activation $\eta(x) = \max\{x, 0\}$ applied element-wise; $\Psi(L, W, S, B)$ is the class of such FNNs with $\max_i\{\|A_i\|_\infty, \|b_i\|_\infty\} \leq B$ and $\sum_{i \leq L} \|A_i\|_0 + \|b_i\|_0 \leq S$. **(ii)** A causal self-attention layer with embedding dimension D and parameters $K, Q, V \in \mathbb{R}^{D \times D}$ maps $X \in \mathbb{R}^{D \times T}$ to $g(X)_i := x_i + V X_{1:i} \mathrm{Softmax}((K X_{1:i})^\top (Q x_i))$, $i \in [T]$, where $X_{1:i} = (x_1, \ldots, x_i)$ and $\mathrm{Softmax}(z) = [e^{z_j} / \sum_{j'} e^{z_{j'}}]_j$, so that each position attends only to itself and the preceding ones; $\mathcal{A}(D, B)$ denotes the class of such layers with $\max\{\|K\|_\infty, \|Q\|_\infty, \|V\|_\infty\} \leq B$. **(iii)** The embedding layer is $\mathrm{Enc}_P(X) = EX + P$ with a learnable $E \in \mathbb{R}^{D \times |\mathcal{V}|}$ and the fixed sinusoidal positional encoding $P = [p_t]_{t=1}^T \in \mathbb{R}^{D \times T}$, $p_{t,2r} = \sin\big(t/10000^{2r/D}\big)$, $p_{t,2r+1} = \cos\big(t/10000^{2r/D}\big)$.

Algorithm 1 Transformer actor–critic with prefix-level exploration

Input: budget n, prompts M, regularization ρ, and the schedule of Section 3.2.
Initialize the score $W_0 \equiv 0$.
for $h = 0, 1, \ldots, H_{\mathrm{tot}} - 1$ **do**
  Actor: $P_h(dy \mid m) \propto \exp\left(W_h(m,y)/\vartheta_h\right) \mu(dy)$,
    i.e., $P_h(\cdot \mid m) = \arg\max_P \{\mathbb{E}_{y\sim P}[W_h(m,y)] - \vartheta_h \mathrm{KL}\left(P \parallel \mu\right)\}$.
  Collect: for $t = 1, \ldots, b_h$: receive $X_t \sim \nu$; generate $Y_t \sim P_h(\cdot \mid X_t)$ prefix by prefix, stopping at the prefix $D_t = \mathsf{Query}_h(X_t, Y_t)$; observe $Z_t = Q(X_t, D_t) + \xi_t$.
  Critic: $\widehat{T}_h \in \arg\min_{T\in\mathcal{T}_h} \frac{1}{b_h} \sum_{t=1}^{b_h} \left[T(X_t, D_t) - Z_t\right]^2$.
  Update: $W_{h+1} = W_h + I_h \cdot \mathrm{clip}_{[-\Delta_h,\Delta_h]}\big(\widehat{T}_h(\cdot, \mathsf{Query}_h(\cdot)) - W_h\big)$ pointwise in (m, y), with the mask $I_h(m,y) \in \{0,1\}$ and the clipping width $\Delta_h(m,y)$.
end for
return: $\widehat{P}(dy \mid m) \propto \exp\left(W_{H_{\mathrm{tot}}}(m,y)/\vartheta_{H_{\mathrm{tot}}}\right) \mu(dy)$.

We define the class of Transformers $\mathcal{T} = \mathcal{T}(N_{\mathrm{blk}}, D, L, W, S, B)$ by

$$\mathcal{T} := \left\{ \mathrm{clip} \circ f_{N_{\mathrm{blk}}} \circ g_{N_{\mathrm{blk}}} \circ \cdots \circ f_1 \circ g_1 \circ \mathrm{Enc}_P \;\middle|\; \begin{array}{l} \|E\|_\infty \le B,\ g_b \in \mathcal{A}(D,B), \\ f_b \in \Psi(L,W,S,B)\ \ (b \in [N_{\mathrm{blk}}]) \end{array} \right\},$$

where N_{blk} is the number of blocks (the depth of the Transformer; a block consists of a self-attention layer g_b followed by an FNN layer f_b), the FNNs are applied column-wise, the last FNN $f_{N_{\mathrm{blk}}}$ has a one-dimensional output, and $\mathrm{clip} = \mathrm{clip}_{[-1,1]}$ is applied element-wise. A network $F \in \mathcal{T}$ thus maps an input sequence to a sequence of scalars in $[-1, 1]$, and the critic's prediction is the entry at the readout position, $F(X)_T$; we identify F with this scalar function. The class is that of Takakura & Suzuki (2023), with causal attention in place of sliding-window attention, a single head, and the output read at the last position.

3.2 Algorithm

Algorithm 1 is an on-policy actor–critic of the kind used in RL post-training. It maintains a *score* $W_h(m,y)$, the current estimate of the reward, and an *actor* P_h, the Gibbs policy of that score at a temperature ϑ_h; since the score is piecewise constant on dyadic cubes, the actor is sampled autoregressively, the probability of the next d bits being the ratio of the Gibbs masses of the children. Each iteration collects a fresh batch of on-policy responses and queries their rewards, fits the critic to them, and updates the actor. Every decision about a sample is a function of its own prefix, the shared critics and the shared rules. What distinguishes the algorithm from a vanilla actor–critic is the choice of the queried prefix.

Depth curriculum and adaptive resolution. Level i of the reward is visible only inside $C_{i-1}(m)$, so the learner has to zoom in along the correct path, while deep bits of a response that is already wrong at a shallow level carry no information **(A4)**. Algorithm 1 therefore uses a *depth curriculum*: at iteration h the actor's confidence is judged at depth h, which increases by one per iteration, and a response that is confident down to that depth is queried ℓ_{look} depths deeper, the *lookahead*. Training on reasoning of increasing length or difficulty is standard practice in RL post-training (Luo et al., 2025; Kimi Team, 2025) and an instance of curriculum learning (Bengio et al., 2009). The critic is a *process* reward model, trained on prefixes of all depths (Lightman et al., 2024; Wang et al., 2024), which allows a response to be stopped early and queried at a coarse prefix whenever the actor is not confident about it (*adaptive resolution*). The query rule is

$$\mathsf{Query}_h(m,y) = \begin{cases} D_k(y) & \text{at the first depth } 1 \le k \le h \text{ with } P_h(D_k(y) \mid m) < \kappa_0, \\ D_{h+\ell_{\mathrm{look}}}(y) & \text{if there is no such depth,} \end{cases} \tag{5}$$

with a constant threshold $\kappa_0 > 0$: generation stops at the first prefix whose actor mass falls below κ_0 and that prefix is queried; if none does, the response is extended and queried ℓ_{look} levels deeper. The rule uses only the probability that the current actor assigns to a prefix, which an autoregressive model already computes while generating it. The lookahead is $\ell_{\mathrm{look}} = 1$ in the small-regularization setting, where the iterations simply continue as long as the budget allows and the output is the last actor, and $\ell_{\mathrm{look}} = R + 1$ with $R = O(\log \mathcal{L})$ and $\mathcal{L} := \log\big(e + n + M + \rho^{-1}\big)$ in the large-budget setting, where they are followed by the step described next.

Training the critic. The critic $\widehat{T}_h$ is the least-squares fit over the entire class $\mathcal{T}_h$ to all records of the batch, including those of the responses that stopped early; all parameters are fitted, and no quantity is computed or memorized per cube. Its risk is controlled by the approximation power and the covering number of $\mathcal{T}_h$, as in the nonparametric analysis of least squares over deep network classes (Schmidt-Hieber, 2020; Takakura & Suzuki, 2023). The batch of iteration h has size $b_h \asymp M^{\zeta_M} c_{h+\ell_{\mathrm{look}}}^{-2}$ up to logarithms, with $\zeta_M := 2 + d/(2\alpha)$: an iteration must reach the accuracy $c_{h+\ell_{\mathrm{look}}}$ of the level it queries (the schedules are given in Appendices D.1 and E.1).

Actor update. Two devices standard in RL post-training are used in the update. In the front phase, the score is replaced only on the cubes of actor mass at least κ_0, where the critic has just been trained on fresh on-policy data, and the mask $I_h(m, y) = \mathbb{1}_{P_h(D_h(y)|m) \geq \kappa_0}$ keeps the old score elsewhere: this is a trust region in the spirit of TRPO and PPO (Schulman et al., 2015; 2017). In the front phase no clipping is needed, $\Delta_h \equiv 2$; the final update is clipped, as in PPO, since a Gibbs policy at temperature ρ amplifies an overestimate of the score exponentially.

Final refinement step (large budgets only). After $H \asymp \rho^{-1/\gamma} + \log(n/\rho)$ exploration iterations the actor has resolved as many levels as ρ can distinguish, and the remaining budget $b_H = n - \sum_{h<H} b_h$ is spent on one last iteration, which samples at temperature ρ, sets $I_H \equiv 1$, and uses

$$\mathsf{Query}_H(m, y) = D_{q(m,y)+\ell_{\mathrm{look}}+r}(y), \qquad \Delta_H(m, y) = \rho + \eta\, c_{1+\lfloor q(m,y)/\bar{\ell} \rfloor}, \tag{6}$$

where $q(m, y)$ is the last exploration iteration at which the region of y was updated, r is a refinement depth and η is a structural constant. Querying r levels deeper than the last update balances approximation against estimation: the reward is Hölder in the local coordinates of each shell $C_{a-1} \setminus C_a$, so a finer query buys a better fit until the estimation error takes over.

3.3 Regret Upper Bound

We give two upper bounds, one for each regime of ρ, proved in Appendices D and E.

Theorem 3. *Suppose that $f \in \mathcal{F}_{M,\theta}$, that the noise is conditionally σ-sub-Gaussian, that $\rho \in [0, 1]$ and that $n \geq M^{\zeta_M}$ with $\zeta_M = 2 + d/(2\alpha)$. Run Algorithm 1 with the hyperparameters of Appendix E.1 in the first case below and of Appendix D.1 in the second; in both cases it uses at most n queries, and its output satisfies, for a fixed exponent $c_{\log} > 0$ and with $\widetilde{O}$ hiding logarithmic factors,*

$$\mathbb{E}\Big[\mathcal{R}_\rho(\widehat{P})\Big] \;\leq\; \begin{cases} \widetilde{O}\Big(\big(n/M^{\zeta_M}\big)^{-\frac{\gamma-1}{2\gamma+1}}\Big) & \text{if } \rho = 0 \text{ or } \rho \lesssim (n/M)^{-\frac{\gamma}{2\gamma+1}}, \\ \widetilde{O}\Big(\rho\big(n\rho^{2+1/\gamma}/M\big)^{-\frac{2\alpha}{2\alpha+d}}\Big) & \text{if } \rho \gtrsim \big(n/M^{\zeta_M}\big)^{-\frac{\gamma}{2\gamma+1}} \mathcal{L}^{c_{\log}}, \end{cases} \tag{7}$$

Theorem 3 leaves a band between its two cases, namely $(n/M)^{-\gamma/(2\gamma+1)} \lesssim \rho \lesssim (n/M^{\zeta_M})^{-\gamma/(2\gamma+1)} \mathcal{L}^{c_{\log}}$, in which the algorithm is well defined but we prove no guarantee. Up to logarithmic factors, the gap between the two thresholds is the price of using a single critic for all prompts: the guaranteed depth is $(n/M^{\zeta_M})^{1/(2\gamma+1)}$ rather than $(n/M)^{1/(2\gamma+1)}$, since estimation theory bounds the L^2 error under the joint query law, not the error on every prefix that a prompt updates (Lemma 23). For a fixed number of prompts the gap is only polylogarithmic, and outside the band both rates are the minimax rates of Section 4 up to logarithmic factors.

Proof sketch. We sketch the proofs through the three inequalities that produce the two rates.

First, the actor keeps exploring the correct path. The two facts support each other by induction on h. If the score is accurate on each cube at the scale of the depth at which that cube was created, then the reward gap between the correct cube of depth h and any cube that left the path exceeds the errors made so far, and the Gibbs actor satisfies $P_h(D_h(y_\infty(m)) \mid m) \geq p_* > 0$ uniformly in h and in the reward (Lemma 12). The correct cube is therefore refined at every iteration, and the fresh batch drawn there restores the accuracy one level deeper; this is what keeps the queries concentrated where the next level is visible, and the same computation shows that the mass leaving the path j levels earlier decays like e^{-cj}, so the queried prefixes stay in a window of logarithmic length and the critic only has to approximate finitely many Hölder functions there.

Second, the depth reached is fixed by the budget. Resolving level i requires an accuracy $\asymp c_i \asymp i^{-\gamma}$ on a cube average, so the batches needed to reach depth H satisfy

$$\sum\nolimits_{h<H} b_h \;\asymp\; M^{\zeta_M} \sum\nolimits_{h<H} c_h^{-2} \;\asymp\; M^{\zeta_M} H^{2\gamma+1} \;\le\; n, \tag{8}$$

up to logarithms, the power ζ_M being the price of converting the risk of the shared critic into accuracy at each updated prefix. In the small-regularization case this is the whole argument: the output is the last actor, uniform on cubes of depth H, its regret is the weight $T_H \asymp H^{-(\gamma-1)}$ of the unresolved levels, and (8) gives $H \asymp (n/M^{\zeta_M})^{1/(2\gamma+1)}$, hence the first exponent of (7). The regularization only adds the penalty $\rho\, dH \log 2$, which stays below T_H in the first case.

Third, in the large-budget case the score has to be turned into a regret bound. A Gibbs policy at temperature ρ cannot distinguish levels of weight below ρ, i.e. levels deeper than $K \asymp \rho^{-1/\gamma}$; reaching that depth costs $M^{\zeta_M}\rho^{-(2+1/\gamma)}$ queries by (8), up to logarithms. What remains is the fine shape of the reward around the correct path, which the last iteration estimates with one more critic. The regret is an integral of variances under the interpolated Gibbs policies. Pointwise accuracy of the held score, clipping and exponential-moment control bound their density ratios relative to the collecting policy, giving schematically

$$\mathbb{E}\big[\mathcal{R}_\rho(\widehat{P})\big] \;\lesssim\; \rho^{-1}\big[(\text{critic error}) + (\text{projection error})\big] + (\text{deep-core and failure errors}) \tag{9}$$

(Proposition 24). Here the projection error is the squared error from replacing the reward by its cube averages; the core and failure terms are lower order. Since the reward is α-Hölder in d local coordinates on each shell, the squared error obeys the nonparametric tradeoff, and balancing it against (8) gives the second case of (7).

4 Minimax Lower Bound

We now show that the rates of Theorem 3 cannot be improved by any algorithm, including algorithms that select prefixes adaptively without any restriction on computation, memory, or the use of past data. Let $\mathcal{N}_\sigma$ be the class of noise distributions that are conditionally mean zero and σ-sub-Gaussian as in Section 2.3, and define the minimax regret

$$R^*(n, M, \rho) := \inf_{\mathcal{A}} \sup_{f\in\mathcal{F}_{M,\theta},\, N\in\mathcal{N}_\sigma} \mathbb{E}_{f,N,\mathcal{A}}\big[\mathcal{R}_\rho(\widehat{P}_{\mathcal{A}})\big], \tag{10}$$

where the infimum is over all learning algorithms that use at most n queries in the protocol of Section 2.3.

Theorem 4. *Fix $\sigma > 0$, suppose that $\lambda \ge \Lambda_{\alpha,d}$ and let $\rho \in [0,1]$. For all $M \ge 1$ and $n \ge M$,*

$$R^*(n, M, \rho) \;\gtrsim\; \begin{cases} (n/M)^{-\frac{\gamma-1}{2\gamma+1}} & \text{if } \rho = 0 \text{ or } \rho \lesssim (n/M)^{-\frac{\gamma}{2\gamma+1}}, \\ \rho \min\Big\{1, \big(n\rho^{2+1/\gamma}/M\big)^{-\frac{2\alpha}{2\alpha+d}}\Big\} & \text{for every } \rho > 0, \end{cases} \tag{11}$$

The proof is given in Appendix F. The condition $\lambda \ge \Lambda_{\alpha,d}$ only ensures that the class $\mathcal{F}_{M,\theta}$ is rich enough to contain the hard instances, with the sufficient constant $\Lambda_{\alpha,d}$ defined in (86) for a bump of amplitude one and radius $1/8$; the bounds hold even if the zoom depths $\ell_i \equiv \ell_\star$ are revealed to the learner.

The table below compares Theorems 3 and 4 regime by regime; all entries are up to polylogarithmic factors. The first cases of (7) and (11) hold under the same condition on ρ, and the second case of (11) holds for every $\rho > 0$. In both regimes the two columns agree in the budget n and in the regularization ρ, and for a fixed number of prompts they agree completely: there the algorithm is *minimax optimal up to logarithmic factors*. Since the lower bounds hold for every algorithm, no amount of computation or memory can improve on our on-policy scheme by more than logarithmic factors. Apart from logarithmic factors, the two columns differ only through M, in the rate when the regularization is small and in the range of ρ when the budget is large; both differences come from converting the joint L^2 error of the shared critic into accuracy at every updated prefix, which makes the cost of reaching depth H in (8) grow as M^{ζ_M} rather than M in the number of prompts.

	Lower bound (Theorem 4)	Algorithm 1 (Theorem 3)
Small regularization	$(n/M)^{-\frac{\gamma-1}{2\gamma+1}}$	$\big(n/M^{\zeta_M}\big)^{-\frac{\gamma-1}{2\gamma+1}}$
Large budget	$\rho\big(n\rho^{2+1/\gamma}/M\big)^{-\frac{2\alpha}{2\alpha+d}}$	$\rho\big(n\rho^{2+1/\gamma}/M\big)^{-\frac{2\alpha}{2\alpha+d}}$

Remark (the two bottlenecks). The two bounds describe two different obstructions. *Small regularization: the budget.* The budget runs out before the levels of weight $c_i > \rho$ are resolved, so the regret is dominated by the weight of the unresolved levels and neither α nor d appears. The bound is the same for $\rho = 0$ and for every $\rho \lesssim (n/M)^{-\gamma/(2\gamma+1)}$, because in that range the regularization changes the optimal value only by $O(\rho^{1-1/\gamma})$, a small fraction of $(n/M)^{-(\gamma-1)/(2\gamma+1)}$. *Large budget: the regularization.* The correct path can be resolved down to the last level of weight $c_i > \rho$, at depth $\rho^{-1/\gamma}$, and what is left is to estimate the fine shape of the reward around it. The second bound holds for every $\rho > 0$, and the first, where it applies, is the larger one.

5 Lower Bound for Fixed-Distribution Sampling: The Benefit of On-Policy Exploration

The algorithm of Section 3 is *on-policy*: the responses whose rewards are queried are generated by the current actor. A common alternative in reward modeling is to collect responses from a fixed base distribution and to fit the reward model offline. In this section, we show that this restriction is costly in the hierarchical setting. Let $\mathcal{A}_{\mathrm{unif}}$ be the class of algorithms that, at every round t, generate the response $Y_t \sim \mu$ uniformly at random (independently of the past), may choose the depth of the queried prefix $D_t \ni Y_t$ after seeing (X_t, Y_t) and the history, and output an arbitrary policy $\widehat{P}$ at the end. The only restriction is the sampling distribution: the estimator, the choice of the query depth and the final policy are unrestricted, and past data may be reused freely. Let $R^*_{\mathrm{unif}}(n, M, \rho)$ be the minimax regret (10) with the infimum restricted to $\mathcal{A}_{\mathrm{unif}}$.

Theorem 5 (The price of fixed sampling)**.** *Suppose that $\sigma > 0$ and $\lambda \geq \Lambda_{\alpha,d}$. There are constants $\rho_0, c_0 > 0$ such that, for all $n \geq M \geq 1$,*

$$R^*_{\mathrm{unif}}(n, M, \rho) \gtrsim \begin{cases} \left(\log \frac{en}{M}\right)^{-(\gamma-1)} & \text{if } \rho \lesssim \left(\log \frac{en}{M}\right)^{-\gamma}, \\ \rho \min\left\{1, \left(n\rho^{2+1/\gamma} e^{-c_0 \rho^{-1/\gamma}}/M\right)^{-\frac{2\alpha}{2\alpha+d}}\right\} & \text{if } \rho \in (0, \rho_0]. \end{cases} \tag{12}$$

The first bound holds for every $\sigma \geq 0$, that is, even for a noiseless oracle.

The proofs are given in Appendix G. The two bounds describe two budget regimes. While $\rho \lesssim (\log(en/M))^{-\gamma}$, the regret of any fixed-sampling method decays at most logarithmically in n/M, whereas in the large-budget regime of Theorem 3, the on-policy actor–critic achieves a polynomial rate; the two conditions are compatible for small ρ. For larger budgets the fixed-sampling regret does decay polynomially, but it reaches that same rate only with the multiplicative price $\exp(c_0\rho^{-1/\gamma})$, whereas the on-policy bound depends on ρ^{-1} polynomially. Two complements are proved there: for a *fixed* $\rho > 0$ and $n \to \infty$ the exponent of n is not worsened by fixed sampling (Proposition 40); and the first bound is tight in its regime (Proposition 39).

Why on-policy sampling escapes. Once the actor has learned that a prefix is promising it generates responses from that prefix, so the cell of the next level is visited with constant probability rather than with probability $2^{-d\ell_\star k}$. Each iteration of Algorithm 1 then resolves one level at a cost of $c_k^{-2} \asymp k^{2\gamma}$ queries instead of $2^{d\ell_\star k}$, which is the statistical benefit of RL post-training over offline reward modeling in this setting.

6 Conclusion

We studied RL post-training with hierarchical rewards, in which the reward of a binary response is a sum of infinitely many Hölder primitives revealed level by level. A Transformer actor–critic of the standard form, alternating on-policy sampling from a KL-regularized policy, least-squares training of a shared Transformer critic and a Gibbs policy update, matches the minimax rates in the budget and in the regularization up to logarithmic factors, and is minimax optimal for a fixed number of prompts. The analysis identifies the role of each component: the prefix mass of the current policy selects where to query, the temperature schedule keeps the correct path confident, the full-class ERM adapts to the unknown cells, and a calibrated, clipped final update transfers the estimation error to the regret. Any method that keeps sampling from the fixed reference distribution μ, in contrast, suffers a logarithmic rate or an exponential price in $\rho^{-1/\gamma}$, which quantifies the benefit of on-policy exploration. Appendix B collects the limitations of our results and the questions they leave open.

AI USE STATEMENT

In this work, generative AI tools assisted with developing the theoretical model and formulating mathematical claims, provided key ingredients for proofs, and assisted with revising mathematical statements, developing and checking proof arguments, and drafting and editing proofs. Additionally, we used generative AI tools for drafting and revising parts of the manuscript, for the code that generates the figure, and for identifying and formatting references. We have reviewed all AI-assisted work. Every theorem, lemma and proof was verified by the authors, and all AI-drafted text was revised by them. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, pp. 41–48, 2009.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Dake Bu, Wei Huang, Andi Han, Atsushi Nitanda, Hau-San Wong, Qingfu Zhang, and Taiji Suzuki. Provable Sample Efficiency of Curriculum Post-Training for Transformer Reasoning. In *International Conference on Machine Learning*, 2026.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep Q-learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 486–489. PMLR, 2020.

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. In *The Thirty Eighth Annual Conference on Learning Theory*, pp. 2026–2142. PMLR, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948v1*, 2025. URL `https://arxiv.org/html/2501.12948v1`.

Satoshi Hayakawa and Taiji Suzuki. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361, 2020.

Juno Kim, Denny Wu, Jason D Lee, and Taiji Suzuki. Metastable dynamics of chain-of-thought reasoning: Provable benefits of search, RL and distillation. In *International Conference on Machine Learning*, 2025.

Kimi Team. Kimi k1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*, 2025.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pp. 681–690, 2008.

Qiyang Li, Yuexiang Zhai, Yi Ma, and Sergey Levine. Understanding the complexity gains of single-task RL with a curriculum. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. DeepScaleR: Surpassing O1-preview with a 1.5B model by scaling RL. Notion blog post, 2025. URL https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2.

Rémi Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Naoki Nishikawa and Taiji Suzuki. Inference-time alignment with rewards in Besov spaces: Provable advantages of feature learning and multi-step policy updates. In *International Conference on Machine Learning*, 2026.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, and Shuiwang Ji. Curriculum reinforcement learning from easy to hard tasks improves LLM reasoning. In *International Conference on Learning Representations*, 2026.

Nived Rajaraman, Audrey Huang, Miro Dudik, Rob Schapire, Dylan Foster, and Akshay Krishnamurthy. Learning to reason with curriculum I: Provable benefits of autocurriculum. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5518–5555. PMLR, 2026. URL https://proceedings.mlr.press/v336/rajaraman26a.html.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. *arXiv preprint arXiv:2410.08146*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zhihong Shao, Yuxiang Luo, Chengda Lu, Z. Z. Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, and Xiaokang Zhang. DeepSeekMath-V2: Towards self-verifiable mathematical reasoning. *arXiv preprint arXiv:2511.22570*, 2025.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33416–33447. PMLR, 2023.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.

Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic simultaneous optimistic optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 19–27. PMLR, 2013.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Siwei Wang, Yifei Shen, Haoran Sun, Shi Feng, Shang-Hua Teng, Li Dong, Yaru Hao, and Wei Chen. Benefits and pitfalls of reinforcement learning for language model planning: A theoretical perspective. In *International Conference on Learning Representations*, 2026a.

Zichen Wang, Haoyang Hong, and Huazheng Wang. When greedy sampling explores: KL-regularized contextual bandits without eluder-dimension dependence. *arXiv preprint arXiv:2609.13564*, 2026b.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I. Jordan. Provably efficient reinforcement learning with kernel and neural function approximations. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

Heyang Zhao, Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Logarithmic regret for online KL-regularized reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77864–77884. PMLR, 2025.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Additional Related Work

This appendix expands the discussion of related work in Section 1.

Prompt-side curricula. Li et al. (2023) show that solving a sequence of easier single-task RL problems, each initialized from the previous solution, reduces the sample complexity of exploring a hard task with sparse rewards. Parashar et al. (2026) schedule reasoning tasks from easy to hard and derive finite-sample bounds within approximate policy iteration. Rajaraman et al. (2026) analyze autocurriculum, in which the model's own performance selects the prompts to train on, and show exponential savings in demonstrations for supervised fine-tuning and a decoupling of the cost of RL fine-tuning from the quality of the reference model. Bu et al. (2026) model chain-of-thought generation as a state-conditioned reasoning tree and prove that, under outcome-only rewards, RL post-training with a depth-increasing or hint-decreasing curriculum has polynomial sample complexity whereas its non-curriculum counterpart faces an exponential bottleneck. In all of these the adaptivity lies in the choice of prompts or tasks and the reasoning space is discrete. Here the prompt distribution is fixed and exogenous, all adaptivity is in the conditional distribution over a continuous response space, and the intermediate signal is not supplied by a curriculum but by the reward itself: the coarse components are informative on their own, so an ordinary on-policy actor–critic localizes the response distribution level by level, with the query depth driven by the mass of the actor rather than by task difficulty or hints. Accordingly, the baseline in our separation is data collection from a fixed response distribution, not RL without a curriculum, and the two forms of adaptivity can be combined.

Search and exploration over reasoning trajectories. Two of the works cited in Section 1 explain the benefit of exploration through mechanisms other than the statistical rate. Kim et al. (2025) model chain-of-thought as a metastable Markov chain in which easy steps form dense clusters and hard steps are sparse transitions between them; search that rewards the sparse edges reduces the expected time to reach new clusters, and the discovered structure can be reinforced by policy gradient or distilled. This is the conceptual prior of our model, namely that exploration finds rare but crucial structure, but the analysis concerns the dynamics of the chain rather than the statistical rate of learning from noisy rewards. Wang et al. (2026a) show on a graph model that RL achieves correct planning through exploration whereas supervised fine-tuning may learn spurious solutions; the result concerns generalization rather than the rate at which the response distribution should be adapted. We ask how fast the reward can be learned when exploration is on-policy, and prove that the answer is minimax optimal and polynomially faster than fixed-distribution sampling.

On-policy exploration under KL regularization. Foster et al. (2025) show, in a linear-softmax framework with a sampling oracle for the pre-trained model, that coverage of the base model is not needed for data efficiency but lower-bounds the runtime of exploration; the focus is the computational–statistical tradeoff. Xiong et al. (2024) formalize KL-regularized RLHF as a contextual bandit with preference feedback and compare offline, online and hybrid data collection; our protocol is an instance of the online case with reward feedback. Zhao et al. (2025) prove that an optimistic algorithm for online KL-regularized bandits with a finite reward class of cardinality N_R has regret $O(\eta\, d_R \log(N_R T))$, η the inverse regularization strength and d_R an eluder-type complexity, because under KL regularization the regret of the Gibbs policy of an estimate is of second order in the estimation error, so that the regret accumulates like a sum of squared errors. We use the same mechanism, in the form of a Bregman-divergence bound of ρ^{-1} times the squared L^2 error of the critic under the actor (Appendix D.3), but for a nonparametric class the squared error decays only at the rate $(n/n_0)^{-2\alpha/(2\alpha+d)}$, so the regret is polynomial in n, and $n_0 = M\rho^{-(2+1/\gamma)}$ queries must first resolve the $\asymp \rho^{-1/\gamma}$ levels that the regularization distinguishes. Wang et al. (2026b) show, for KL-regularized bandits, that greedily sampling from the Gibbs policy of an estimated reward already explores, with polylogarithmic regret and no explicit eluder-dimension dependence when the regularization is strong enough. The observation that a Gibbs actor explores by itself is therefore not new and we do not claim it; our actor is the Gibbs policy of a held score with a temperature schedule (ϑ_h) and a final step at temperature ρ (Section 3.2), and our contribution is what this exploration achieves under hierarchical nonparametric rewards, where no eluder-type structure is available: the minimax rates of Theorem 4 and the polynomial separation of Section 5.

Continuum-armed bandits. The zooming algorithm (Kleinberg et al., 2008) refines the discretization of a Lipschitz bandit near the arms that appear near-optimal, and HOO (Bubeck et al., 2011, Theorem 6) samples a hierarchical partition optimistically, refining the cells around the maximizer and keeping the rest coarse, with regret $\widetilde{O}(n^{(d'+1)/(d'+2)})$ for any d' larger than the near-optimality dimension. For $f(x) = 1 - \|x - x^*\|^a$, $a \geq 1$, the matched dissimilarity $\|x - y\|^a$ gives near-optimality dimension zero (Bubeck et al., 2011, Example 3). The corresponding local-HOO guarantee is $\widetilde{O}(\sqrt{n})$ (Bubeck et al., 2011, Example 4 and Theorem 10). SOO (Munos, 2011) and its noisy version StoSOO (Valko et al., 2013) remove the knowledge of the smoothness by expanding several depths simultaneously, assuming local smoothness only around the global maximizer. Our algorithm shares the intuition of coarse-to-fine partitioning with queries concentrated where the reward is high, and differs in three respects. First, the near-optimality dimension governs the rate because the diameters of the cells shrink geometrically with depth; in our hierarchy the gap of a response leaving the correct path at level i is $\asymp i^{-(\gamma-1)}$ while its distance to the maximizer is $\asymp 2^{-\ell_\star i}$, so the gap decays only polynomially in the depth and the model falls outside the weak-Lipschitz framework. This produces the exponent $(\gamma - 1)/(2\gamma + 1)$ of the small-regularization bound of Theorem 3, which involves neither α nor d: it reflects the cost $\asymp i^{2\gamma}$ of resolving level i from the noise, not the packing of near-optimal sets. Second, the KL regularization makes the target the Gibbs policy P_ρ^* rather than the maximizer, and the large-budget rate $\rho\,(n/n_0)^{-2\alpha/(2\alpha+d)}$ is a nonparametric estimation rate of the reward on the shells around the maximizer, which does not arise when only the maximizer matters. Third, motivated by LLM reasoning with several successive bottlenecks and by the way LLMs are trained, the optimal rates are attained by an actor–critic that fits a causal Transformer by ERM on fresh on-policy data and updates the actor as a Gibbs policy, not by optimistic tree search with explicit confidence bounds.

Neural function approximation in estimation and RL. Deep networks attain minimax rates for Hölder and Besov classes and over sparse parameter spaces (Schmidt-Hieber, 2020; Suzuki, 2019; Hayakawa & Suzuki, 2020), and Takakura & Suzuki (2023) give approximation and estimation bounds for Transformers with sequence inputs, the class we use for the critic (Section 3.1); our covering-number bound (Appendix C.4) adapts their argument to the causal architecture and to prefix queries. Fan et al. (2020) propagate the nonparametric regression error of deep ReLU networks through fitted Q-iteration, and Yang et al. (2020) bound the regret of optimistic value iteration with kernel and overparameterized neural approximation by an effective dimension. Both share our position that the statistical cost is governed by the complexity of the neural class, but they estimate value functions of a fixed MDP with exploration by optimism bonuses (or a fixed data distribution), whereas we estimate a reward on a continuous response space, exploration arises from the mass of the Gibbs actor on promising prefixes, and the rates are matched by lower bounds.

B Limitations and Future Work

Our upper bounds are not optimal in the number of prompts M, for the reason given in Section 4; removing the extra factor, for instance through a prompt-wise estimation guarantee or an analysis of the exploration in expectation over the prompts, is left open. The two upper bounds are also separated by the band described after Theorem 3, where the same algorithm is well defined but we prove no guarantee, and the two parameter settings differ in the lookahead and in the presence of the refinement step. We assumed exact ERM and exact Gibbs sampling; the optimization remains to be analyzed. An empirical study of the prefix-mass exploration rule on synthetic hierarchical rewards and reasoning benchmarks is left for future work.

C Common Preliminaries and Tools

This section collects the material that is used by more than one of the proofs that follow: the conventions and notation of the whole appendix (Appendix C.1), the geometry of hierarchically nested rewards (Appendix C.2), the Gibbs mass of the correct cube under a birth-accurate score (Appendix C.3), and the approximation and estimation theory of the Transformer critic on the relevant window (Appendix C.4). The geometry of Appendix C.2 is used by the upper bounds of Appendices D and E and again by the lower bounds of Appendices F and G; the remaining subsec-

tions are common to the two upper bounds. The results specific to one theorem, together with the hyperparameter schedules that they use, are stated in the section devoted to that theorem.

C.1 Conventions and Notation

α-Hölder smoothness. For $\alpha > 0$, put $k = \lceil \alpha \rceil - 1$, $\tau = \alpha - k \in (0, 1]$ and $\beta = \min\{\alpha, 1\}$. For $v \in C^k(\mathbb{R}^d)$ define

$$|v|_\alpha := \max\left\{ \sum_{1 \le |s| \le k} \|D^s v\|_\infty,\ \max_{|s|=k} [D^s v]_\tau \right\}, \qquad [w]_\tau := \sup_{u \ne v} \frac{|w(u) - w(v)|}{\|u - v\|_\infty^\tau}, \tag{13}$$

where s is a multi-index and the empty sum is zero. On a cube, $\|v\|_{C^\alpha} := \|v\|_\infty + |v|_\alpha$ uses the same restricted norms. We call $v : \mathbb{R}^d \to \mathbb{R}$ *α-Hölder smooth with constant* λ if $v \in C^k(\mathbb{R}^d)$ and $|v|_\alpha \le \lambda$. Thus integer smoothness means that derivatives of order $\alpha - 1$ are Lipschitz; for $\alpha \le 1$, $|v|_\alpha = [v]_\alpha$. Condition **(A1)** preserves the original Hölder condition for $\alpha \le 1$ and implies $|g_i^0(u) - g_i^0(v)| \le \lambda \|u - v\|_\infty^\beta$ for every $\alpha > 0$: when $\alpha > 1$, integrate the gradient along the segment and use $\sum_j \|\partial_j g_i^0\|_\infty \le \lambda$. We use C^α (or α-Hölder) for the full smoothness of (13), and β for function-value oscillations.

Conventions. Throughout the appendix, $C, C', c, c', \dots$ denote positive constants that depend only on the structural parameters $\theta = (d, \alpha, \lambda, c_{\min}, \gamma, \ell_\star, \bar{\ell})$ and on σ, and never on n, M, ρ; their values may change from line to line. We fix a reward $f \in \mathcal{F}_{M,\theta}$ and suppress the prompt index m when a single prompt is considered. We use the following notation:

$$c_i = \frac{i^{-\gamma}}{\zeta(\gamma)}, \qquad T_k = \sum_{i > k} c_i, \qquad L_i = \sum_{a \le i} \ell_a, \qquad E_a(m) = C_{a-1}(m) \setminus C_a(m),$$

$$f^{[k]} = \sum_{i \le k} c_i h_i, \qquad G_m = \max_y f(y; m), \qquad r_\star = 2^{-\beta \ell_\star}, \qquad \chi = 3^\gamma r_\star, \qquad B_0 = 2^{d\bar{\ell}},$$

$$\kappa = \frac{c_{\min}}{2}, \qquad K = \rho^{-1/\gamma}, \qquad \mathcal{L} = \log\bigl(e + n + M + \rho^{-1}\bigr),$$

$$p = \frac{2\alpha}{2\alpha + d}, \qquad p_0 = \frac{\gamma - 1}{2\gamma + 1}.$$

The abbreviations p and p_0 for the two exponents are used in the appendix only (they are recalled where needed). Both upper bounds measure the accuracy of a score at the depth of the cube that carries it, and both pay a power of M in the batch sizes; we write

$$a_j = c_{1 + \lfloor j/\bar{\ell} \rfloor} \ \ (j \ge 0), \qquad \zeta_M = 2 + \frac{d}{2\alpha}, \qquad \ell_N = 1 + \log(N + 1), \tag{14}$$

for the known scale attached to a depth j (so that $a_j \asymp c_{k(j)} \asymp (j+1)^{-\gamma}$ for $k(j) = \min\{i : j \le L_i\}$, by **(A3)**), the exponent of M in the budget, and a logarithmic factor of an approximation size N.

Table 1 collects the symbols of this list that recur throughout the appendix, and Table 2 the quantities that the two upper bounds share. The schedules that belong to one proof only are listed in Table 3 for the large-budget bound of Theorem 3 and in Appendix E.1 for the small-regularization bound of Theorem 3.

By **(A3)**,

$$\chi \le \frac{1}{2}, \qquad \lambda\chi \le \frac{c_{\min}}{8}, \tag{15}$$

and for $k \ge 1$,

$$c_T k^{-(\gamma-1)} \le T_k \le C_T k^{-(\gamma-1)}, \qquad c_T = \frac{2^{-(\gamma-1)}}{(\gamma-1)\zeta(\gamma)}, \quad C_T = \frac{1}{(\gamma-1)\zeta(\gamma)}. \tag{16}$$

For a cube D we write $Q(D) = Q(m, D)$ for the cube average (4); the cube average always carries its arguments, and the cubes $Q_0, Q_1, Q_\flat$ of the lower-bound constructions never do. The letters

Table 1: Notation used throughout the appendix.

Symbol	Meaning
$c_i = i^{-\gamma}/\zeta(\gamma)$	weight of level i of the hierarchy
$T_k = \sum_{i>k} c_i$	weight of the levels deeper than k
$L_i = \sum_{a\le i} \ell_a$	depth at which level i is fully resolved
$C_i(m)$, $E_a(m)$	active cell of level i; shell $C_{a-1} \setminus C_a$
$f^{[k]} = \sum_{i\le k} c_i h_i$	reward truncated at level k
$G_m = \max_y f(y;m)$	optimal reward of prompt m
$Q(m, D)$	cube average (4) queried by the oracle
$\kappa = c_{\min}/2$	gap constant of Lemma 8
$r_\star$, χ, B_0	zoom contraction, its γ-weighted version, branching
$p = \frac{2\alpha}{2\alpha+d}$, $p_0 = \frac{\gamma-1}{2\gamma+1}$	the two exponents of Theorem 3

Table 2: Notation shared by the two upper bounds.

Symbol	Meaning	Defined in
$a_j = c_{1+\lfloor j/\bar{\ell}\rfloor}$	accuracy scale attached to depth j	(14)
$\zeta_M = 2 + \frac{d}{2\alpha}$	exponent of M in the batch sizes	(14)
W_h, $\mathcal{D}_h(m)$	score at phase h; partition on which it is constant	Lemma 19
η	birth-accuracy level	Definition 11
$P_h^*(m)$, $k(h)$	correct depth-h cube; the level it has reached	Appendix C.3
p_*	Gibbs mass of the correct cube	Lemma 12
ϑ_h, κ_0	temperature schedule; mass threshold	Appendix C.3
S	lookahead: query depth minus phase	Section 3.2
H	number of front-phase iterations	Appendices D.1 and E.1
$\varepsilon_h = \frac{\eta}{4} a_{h+S}$	accuracy required of the critic of phase h	(52)
$\mathcal{W}_h$, K_w, J, N	query window; its size, length, approximation size	Lemma 17
p_S	actor mass of a single fine query	Lemma 21
b_h, δ_h	batch size and failure budget of phase h	Appendices D.1 and E.1
$\mathcal{E}$	event that all front-phase critics are accurate	(52)

D, L, W, S, T, B that name the parameters of the Transformer class in Section 3.1 all denote other objects in this appendix (D is a cube, L_i a depth, W_h a score, S the lookahead, T_k a tail weight and B_0 a branching number), so the class is written here as

$$\mathcal{T}(N_{\mathrm{blk}}, D_{\mathrm{emb}}, L_{\mathrm{FFN}}, W_{\mathrm{FFN}}, S_{\mathrm{FFN}}, B_{\mathrm{par}}) \quad \text{on inputs of length } T_{\mathrm{in}},$$

where D_{emb} is the embedding dimension, $L_{\mathrm{FFN}}, W_{\mathrm{FFN}}, S_{\mathrm{FFN}}$ the depth, width and sparsity of the FNN layers (also called FFN layers here), B_{par} the bound on the parameters and T_{in} the input length; N_{blk} is the number of blocks, as in the main text. Boundaries of dyadic cubes are μ-null sets and are assigned by the fixed convention of Section 2.1; all statements below are understood up to such null sets.

C.2 Geometry of Hierarchical Rewards

The loss of a response is comparable to the tail weight of the level at which it leaves the correct path: (17) makes exploration necessary and (18) makes reaching a deep cell sufficient.

Lemma 6 (Basic geometry)**.** *Under Assumption 2, for every m, the series* (1) *converges uniformly, $f(\cdot;m) \in [-1,1]$, and $\bigcap_{i\ge 0} C_i(m) = \{y_\infty(m)\}$ is the unique maximizer of $f(\cdot;m)$. Moreover, with $C_{\mathrm{near}} = 2 + \lambda 2^\gamma r_\star/(1-\chi)$,*

$$G_m - f(y;m) \ge \kappa T_i \qquad (y \in E_i(m)), \tag{17}$$

$$0 \le G_m - f(y;m) \le C_{\mathrm{near}} T_k \qquad (y \in C_k(m)). \tag{18}$$

Proof. Uniform convergence and $f \in [-1,1]$ follow from $|h_i| \le 1$ and $\sum_i c_i = 1$.

Convergence and boundedness are immediate from $|h_i| \le 1$ and $\sum_i c_i = 1$. For the two gaps we compare an arbitrary response with the maximizer: the levels above the exit contribute nothing, the levels below it contribute at least a fixed fraction of their weight by **(A2)**, and the levels already resolved contribute an error that the zoom condition **(A3)** keeps below a fraction of the gain; summing the resulting geometric series gives (17), and the same comparison inside a cell gives (18).

Lower bound on the loss outside the cells. Let $z = y_\infty(m) \in \bigcap_i C_i(m)$ and $y \in E_a(m)$. For every i we have $z \in C_i(m)$, i.e., z lies in the cube of relative depth ℓ_i that contains the maximizer of $g_{i,m}$, so by **(A1)**, **(A2)** and (15),

$$h_i(z;m) \ge g_{i,m}(z^\star_{i,m}) - \lambda r_\star \ge c_{\min} - \frac{c_{\min}}{8} = \frac{7c_{\min}}{8} \qquad (i \ge 1); \tag{19}$$

the same bound holds for $i \le k$ and any $z \in C_k(m)$, which is the form used in Lemma 8. For $i > a$ we have $h_i(y;m) = 0$. For $i < a$, both y and z lie in $C_{a-1}(m)$, which is a cube of relative side at most $2^{-\ell_\star(a-i)}$ in the coordinates of $g_{i,m}$, so that

$$|h_i(z;m) - h_i(y;m)| \le \lambda r_\star^{a-i} \qquad (i < a),$$

while for $i = a$ we have $h_a(y;m) \le g^\star_{a,m} \le h_a(z;m) + \lambda r_\star$. Since $(a+1)/i \le 3^{a-i}$ for $i < a$, we obtain

$$\lambda\left(c_a r_\star + \sum_{i<a} c_i r_\star^{a-i}\right) \le \lambda c_{a+1}\left(2^\gamma r_\star + \sum_{b\ge 1} 3^{\gamma b} r_\star^b\right) = \lambda c_{a+1}\left(2^\gamma r_\star + \frac{\chi}{1-\chi}\right) \le \frac{3c_{\min}}{8} c_{a+1}. \tag{20}$$

Combining (19) and (20),

$$f(z;m) - f(y;m) \ge \frac{7c_{\min}}{8} \sum_{i=a+1}^{\infty} c_i - \frac{3c_{\min}}{8} c_{a+1} \ge \frac{c_{\min}}{2} T_a = \kappa T_a.$$

Since $G_m \ge f(z;m)$, this gives (17), and shows that no $y \notin \bigcap_i C_i(m)$ is a maximizer.

Upper bound on the loss inside the cells. If $y \in C_k(m)$ and $z \in C_{k'}(m)$ with $k' \ge k$, then by the same geometric series and $|h_i| \le 1$,

$$f(z;m) - f(y;m) \le \sum_{i\le k} c_i \lambda r_\star^{k-i+1} + 2T_k \le C_{\text{near}} T_k,$$

which gives (18); the maximum is attained at the common point of the cells by continuity. □

A query at depth h sees only the first $\max\{1, \lceil h/\ell_\star \rceil\}$ levels: a deeper primitive lies either inside the queried cube, where it has mean zero **(A4)**, or outside it.

Lemma 7 (Cube averages resolve only finitely many levels)**.** *Let D be a dyadic cube of depth h and $K_h = \max\{1, \lceil h/\ell_\star \rceil\}$. Then*

$$Q(m, D) = \frac{1}{\mu(D)} \int_D f^{[K_h]}(y;m)\, \mathrm{d}\mu\,(y). \tag{21}$$

More generally, for any $i \ge 2$ with $L_{i-1} \ge h$, either $C_{i-1}(m) \subset D$ or the interiors of $C_{i-1}(m)$ and D are disjoint, and in both cases $\int_D h_i(\cdot;m)\, \mathrm{d}\mu = 0$.

Proof. If $L_{i-1} \ge h$, then $C_{i-1}(m)$ is at least as deep as D; since two dyadic cubes are either nested or have disjoint interiors, it is contained in D or disjoint from it. In the former case, by **(A4)**,

$$\int_D h_i(\cdot;m)\, \mathrm{d}\mu = \mu(C_{i-1}(m)) \int_\Omega g_{i,m}\, \mathrm{d}\mu = 0.$$

Since $L_{i-1} \ge \ell_\star(i-1) \ge h$ for $i > K_h$, (21) follows. □

Leaving the correct path at level a costs the whole weight $\sum_{i>a} c_i$ of the levels below a, up to the resolution c_k of the current depth.

Lemma 8 (Gap between the correct cube and the others)**.** *Fix m and let $\mathcal{D}$ be a finite partition of Ω into dyadic cubes. Let $D_* \in \mathcal{D}$ be the cube containing $y_\infty(m)$, let $j \geq 1$ be its depth and*

$$k = \min\{i \geq 1 : j \leq L_i\}, \qquad \textit{so that} \qquad C_k(m) \subset D_* \subset C_{k-1}(m).$$

Then every $D \in \mathcal{D} \setminus \{D_*\}$ satisfies either $D \subset E_a(m)$ for a unique $a \leq k-1$, or $D \subset C_{k-1}(m) \setminus D_*$, and

$$Q(D_*) - Q(D) \geq \kappa \sum_{i=a+1}^{k} c_i - 3c_k \qquad (D \subset E_a(m),\ a < k), \tag{22}$$

$$Q(D_*) - Q(D) \geq -3c_k \qquad (D \subset C_{k-1}(m) \setminus D_*). \tag{23}$$

Moreover,

$$\sup_{u,v \in C_{k-1}(m)} \left| f^{[k]}(u;m) - f^{[k]}(v;m) \right| \leq 3c_k. \tag{24}$$

Proof. If $D \neq D_*$ contained $C_a(m)$ for some $a \leq k-1$, it would contain D_*, contradicting disjointness; hence, for each $a \leq k-1$, D is inside $C_a(m)$ or disjoint from it, which gives the classification. By Lemma 7, $Q(D_*)$ is the average of $f^{[k]}$ over D_*; moreover $f = f^{[a]} = f^{[k]}$ on $E_a(m)$ and $f = f^{[k]}$ on $C_{k-1}(m) \setminus C_k(m)$. The proof of Lemma 6 gives, for $z \in C_k(m)$ and $y \in E_a(m)$,

$$f^{[k]}(z;m) - f^{[k]}(y;m) \geq \kappa \sum_{i=a+1}^{k} c_i,$$

and, for $u, v \in C_{k-1}(m)$,

$$\left| f^{[k]}(u;m) - f^{[k]}(v;m) \right| \leq 2c_k + \lambda \sum_{i<k} c_i r_\star^{k-i} \leq 3c_k,$$

which is (24). Since $Q(D_*)$ is an average over a subset of $C_{k-1}(m)$, (24) gives $Q(D_*) \geq f^{[k]}(z;m) - 3c_k$ for every $z \in C_k(m)$, and averaging the first display over $y \in D$ yields (22) and (23). □

Inside a cube that the hierarchy has not yet split, cube averages at a fixed depth differ by at most the weight of the current level; (26) is the pointwise version on a shell.

Lemma 9 (Oscillation of cube averages at finite depth)**.** *Let A be a dyadic cube of depth h, let $k(h) = \min\{i \geq 1 : h \leq L_i\}$ (so that $L_{k(h)-1} < h \leq L_{k(h)}$ for $h \geq 1$), and let $R \geq 0$. For all dyadic cubes $D, D' \subset A$ of depth at most $h + R + 1$,*

$$|Q(D) - Q(D')| \leq C(R+1)c_{k(h)}. \tag{25}$$

Moreover, if $A \subset E_i(m)$ for some i, then for every dyadic cube $D \subset A$ and $y \in D$,

$$|f(y;m) - Q(D)| \leq C_{\mathrm{osc}} c_i. \tag{26}$$

Proof. A level i with $L_{i-1} < h$ varies on A by at most $\lambda c_i 2^{-\beta(h - L_{i-1})}$ by **(A1)**. Since $c_i/c_{k(h)} \leq (k(h) - i + 1)^\gamma$ and $h - L_{i-1} \geq \ell_\star(k(h) - i)$, the sum over such levels is at most

$$\lambda c_{k(h)} \sum_{r' \geq 0} (r'+1)^\gamma 2^{-\beta \ell_\star r'} \leq C c_{k(h)}.$$

Levels with $h \leq L_{i-1} < h + R + 1$ number at most $(R+1)/\ell_\star + 1$, and each has amplitude at most $c_i \leq c_{k(h)}$. Deeper levels do not contribute to $Q(D)$ by Lemma 7. This proves (25); (26) follows in the same way, using that on $E_i(m)$ only the levels $\leq i$ are active and $c_a/c_i \leq (i - a + 1)^\gamma$ for $a \leq i$. □

On a shell the reward is a finite sum of primitives, hence α-Hölder in the d local coordinates of that shell, with a seminorm of the order of a_j.

Lemma 10 (The reward on a shell is Hölder in local coordinates)**.** *Let $P_j^* = b_j(m) + 2^{-j}\Omega$ be the depth-j cube on the correct path, $A_j = P_j^* \setminus P_{j+1}^*$, and $i(j) = \min\{i : L_i > j\}$. On A_j, $f = f^{[i(j)]}$, and the function*

$$g_{m,j}(z) = f^{[i(j)]}\big(b_j(m) + 2^{-j}z; m\big), \qquad z \in \mathbb{R}^d,$$

satisfies $|g_{m,j}| \le 1$ and

$$\|g_{m,j} - g_{m,j}(0)\|_{C^\alpha(\Omega)} + [g_{m,j}]_\beta \le C c_{i(j)} \le C a_j, \tag{27}$$

with a_j as in (14)*.*

Proof. Levels $i > i(j)$ are supported in P_{j+1}^*, hence vanish on A_j. Put $t_a = 2^{-(j-L_{a-1})} \le 1$ for $a \le i(j)$. By **(A1)**, a derivative of order $1 \le |s| \le k$ of the rescaled primitive has sup norm at most $\lambda t_a^{|s|}$, its order-k derivative has τ-Hölder seminorm at most λt_a^α, and its function-value β-Hölder seminorm is at most λt_a^β. Each exponent is at least β, and subtraction of its value at zero also bounds its sup norm on Ω by λt_a^β. Thus, by the same geometric series as in Lemma 9,

$$\|g_{m,j} - g_{m,j}(0)\|_{C^\alpha(\Omega)} + [g_{m,j}]_\beta \le C\lambda \sum_{a \le i(j)} c_a 2^{-\beta(j-L_{a-1})} \le C c_{i(j)}.$$

Also, $c_{i(j)} \le C a_j$ because the indices $i(j) \le 1 + \lfloor j/\ell_\star \rfloor$ and $1 + \lfloor j/\bar{\ell} \rfloor$ differ by at most the bounded factor $\lceil 2\bar{\ell}/\ell_\star \rceil + 1$. □

C.3 Gibbs Mass of the Correct Cube under Birth-Accurate Scores

Throughout, $P_h^*(m) = D_h(y_\infty(m))$ denotes the correct depth-h cube and $k(h) = \min\{i \ge 1 : h \le L_i\}$, so that $C_{k(h)}(m) \subset P_h^*(m) \subset C_{k(h)-1}(m)$; we write $k(0) = 1$. The scores of Algorithm 1 are analyzed through the following notion. A cube created early is only ever estimated at the coarse scale of its own depth, so the accuracy the analysis can require of a score is scale-dependent.

Definition 11 (Birth-accurate score)**.** Let $\mathcal{D}$ be a finite partition of Ω into dyadic cubes and let W be constant on the cubes of $\mathcal{D}$. We say that W is *η-birth-accurate on $\mathcal{D}$* if

$$|W(D) - Q(m, D)| \le \frac{\eta}{4}\, a_{\mathrm{depth}(D)} \qquad \text{for all } D \in \mathcal{D}. \tag{28}$$

The accuracy allowed on a cube is tied to the depth of the cube, i.e., to the scale of the level that a query at this depth resolves, and not to a common final accuracy.

Lemma 12 (Gibbs mass and cold tail)**.** *Fix a prompt m and a phase $h \ge 1$, and let $\mathcal{D}_h$ be the dyadic leaf partition on which the current score W_h is constant, assumed to be such that $P_h^*(m)$ is a union of cubes of $\mathcal{D}_h$. Let $k = k(h)$ and suppose that W_h is η-birth-accurate on $\mathcal{D}_h$ in the sense of Definition 11. Suppose that*

$$0 < t_- \le \frac{\vartheta}{c_k} \le t_+, \qquad B_0 \exp\left(-\frac{\kappa}{2t_+}\right) \le \frac{1}{2}, \qquad \eta \le \eta_0, \tag{29}$$

where η_0 is a structural constant. Then, uniformly in the partition, in h and in the zoom sequence, the actor $P_h = P_{W_h,\vartheta}$ of Algorithm 1, line 3, satisfies

$$P_{W_h,\vartheta}(P_h^*(m) \mid m) \ge p_* := \Big[1 + 2B_0 e^{C_{\mathrm{gap}}/t_-}\Big]^{-1}, \tag{30}$$

$$P_{W_h,\vartheta}\big(\big(P_{h-j}^*(m)\big)^c \,\big|\, m\big) \le C e^{-cj} \qquad (0 \le j \le h), \tag{31}$$

with a structural constant C_{gap}. For $h = 0$, $P_0^* = \Omega$ has mass one.

Proof. The proof compares the Gibbs weight inside the correct cube with the total Gibbs weight outside it, through the identity

$$\frac{1}{P_{W_h,\vartheta}(P_h^* \mid m)} = 1 + \frac{\int_{\Omega \setminus P_h^*} e^{W_h/\vartheta}\, \mathrm{d}\mu}{\int_{P_h^*} e^{W_h/\vartheta}\, \mathrm{d}\mu}.$$

The step *Inside* lower-bounds the denominator by Jensen's inequality, and the step *Outside* upper-bounds the numerator shell by shell, using the reward gap of Lemma 8 and the birth accuracy of the score; combining the two gives the constant lower bound (30), and summing the same geometric series over the deep shells only gives the cold tail (31).

Inside: lower bound on the Gibbs weight of the correct cube. The cubes $D \in \mathcal{D}_h$ inside P_h^* have depth at least h, so $a_{\mathrm{depth}(D)} \le a_h \le Cc_k$ and $W_h(D) \ge Q(D) - C\eta c_k$ on them. Since $Q(P_h^*)$ is the volume-weighted average of the $Q(D)$, $D \subset P_h^*$ (tower identity), Jensen's inequality gives

$$\int_{P_h^*} e^{W_h/\vartheta}\,\mathrm{d}\mu = \sum_{D\subset P_h^*} \mu(D) e^{W_h(D)/\vartheta} \ge \mu(P_h^*)\exp\left(\frac{Q(P_h^*) - C\eta c_k}{\vartheta}\right); \tag{32}$$

no constancy of W_h inside P_h^* is needed.

Outside: upper bound on the Gibbs weight of the outer shells. Apply Lemma 8 to the partition $\{P_h^*\} \cup \{D \in \mathcal{D}_h : D \cap P_h^* = \emptyset\}$, in which P_h^* is the cube containing $y_\infty(m)$: every $D \in \mathcal{D}_h$ disjoint from P_h^* satisfies either $D \subset E_a(m)$ for a unique $a < k$, or $D \subset C_{k-1}(m) \setminus P_h^*$, and (22), (23) hold with $D_* = P_h^*$. A cube $D \subset E_a(m)$ is strictly inside $C_{a-1}(m)$, so $\mathrm{depth}(D) \ge L_{a-1}+1$ and $a_{\mathrm{depth}(D)} \le Cc_a$ by **(A3)**; similarly $a_{\mathrm{depth}(D)} \le Cc_k$ for $D \subset C_{k-1} \setminus P_h^*$. Hence, by (28), (22) and $c_a \le 2^\gamma c_{a+1} \le 2^\gamma \sum_{i=a+1}^k c_i$, for $D \subset E_a(m)$,

$$\begin{aligned} Q(P_h^*) - C\eta c_k - W_h(D) &\ge \kappa \sum_{i=a+1}^{k} c_i - 3c_k - C\eta(c_a + c_k) \\ &\ge \frac{\kappa}{2}\sum_{i=a+1}^{k} c_i - C_{\mathrm{gap}} c_k \ge \Big[\frac{\kappa}{2}(k-a) - C_{\mathrm{gap}}\Big] c_k, \end{aligned}$$

once $\eta \le \eta_0$ with η_0 small (the old errors are absorbed into half of the gap), and the same bound with the right-hand side $-C_{\mathrm{gap}}c_k$ holds for $D \subset C_{k-1} \setminus P_h^*$. By $h \le L_k$ and **(A3)**, the volumes satisfy $\mu(E_a)/\mu(P_h^*) \le B_0^{k-a+1}$ and $\mu(C_{k-1})/\mu(P_h^*) \le B_0$. Summing the volumes of the cubes in each shell and using (32) and (29),

$$\begin{aligned} \frac{1}{P_{W_h,\vartheta}(P_h^* \mid m)} &\le 1 + \sum_{D\cap P_h^* = \emptyset} \frac{\mu(D)}{\mu(P_h^*)} \exp\left(-\frac{Q(P_h^*) - C\eta c_k - W_h(D)}{\vartheta}\right) \\ &\le 1 + B_0 e^{C_{\mathrm{gap}}/t_-}\left[1 + \sum_{b'=1}^{k-1}\left(B_0 e^{-\kappa/(2t_+)}\right)^{b'}\right] \\ &\le 1 + 2B_0 e^{C_{\mathrm{gap}}/t_-}, \end{aligned}$$

which is (30). For (31), a point outside P_{h-j}^* lies in a shell E_a with $L_{a-1} < h - j$, hence $L_a < h - j + \bar{\ell}$ and $k - a \ge j/\bar{\ell} - 1$; summing the same geometric series only over these shells gives $P_{W_h,\vartheta}((P_{h-j}^*)^c \mid m) \le C\sum_{b' \ge j/\bar{\ell}-1} 2^{-b'} \le Ce^{-cj}$. □

The temperature schedule satisfies (29)**.** Both hyperparameter settings use the same schedule $\vartheta_h = t_0\, c_{1+\lfloor h/\ell_\star \rfloor}$ (Appendices D.1 and E.1), and we check here, once for both, that it meets the condition (29) of Lemma 12. Let $D_0 = \lceil 2\bar{\ell}/\ell_\star \rceil + 1$. If depth h corresponds to the level $k(h)$, then $1 + \lfloor h/\ell_\star \rfloor$ and $k(h)$ differ by at most a factor D_0, so

$$\frac{\vartheta_h}{c_{k(h)}} \in \big[t_0 2^{-\gamma} D_0^{-\gamma},\ 2^\gamma t_0\big]. \tag{33}$$

We fix t_0 so small that $B_0 \exp\big(-\kappa/(2^{\gamma+1}t_0)\big) \le 1/2$; then Lemma 12 applies at every step with $t_+ = 2^\gamma t_0$ and $t_- = t_0 2^{-\gamma} D_0^{-\gamma}$, and p_* is a structural constant (independent of S, n, M, ρ).

C.4 Window-Restricted Transformer Approximation and Estimation

The critic is fitted on data collected by the actor, and the two upper bounds both need the same chain of arguments to turn this into a pointwise accuracy statement:

windowed target decomposition → Transformer realization → covering number → risk.

Lemma 13 shows that a causal Transformer can read the token positions that identify a queried cube; Lemma 14 bounds the covering number of the full class; Lemma 15 converts an approximation error and a covering number into a bound on the conditional risk of the empirical minimizer; Lemma 16 shows that, on the window that the actor actually queries, the infinite hierarchical target reduces to finitely many d-dimensional Hölder functions; and Lemma 17 realizes approximations of those functions inside the Transformer class. How the resulting risk bound is turned into accuracy at every updated node is not a generic estimation question and is treated where it is used (Lemma 23).

Position separation by the sinusoidal encoding. Take D_{emb} to be a multiple of 16. The coordinates $r' = 0$ and $r' = D_{\mathrm{emb}}/16$ of the encoding contain

$$\varphi(t) = \Big(\sin t,\ \cos t,\ \sin\Big(t/\sqrt{10}\Big),\ \cos\Big(t/\sqrt{10}\Big)\Big), \tag{34}$$

since $10000^{2(D_{\mathrm{emb}}/16)/D_{\mathrm{emb}}} = \sqrt{10}$. The constructions below read the bits of a queried cube by attention, which requires the fixed sinusoidal encoding to separate positions well enough.

Lemma 13 (Position separation)**.** *For integers $1 \le r' \le T_{\mathrm{in}}$,*

$$\Gamma(r') := (1 - \cos r') + \Big(1 - \cos\Big(r'/\sqrt{10}\Big)\Big) \ge \frac{1}{40T_{\mathrm{in}}^2}. \tag{35}$$

Consequently, for integer positions $s, t \in [1, T_{\mathrm{in}}]$,

$$\langle\varphi(s), \varphi(t)\rangle = 2 \ \ (s = t), \qquad \langle\varphi(s), \varphi(t)\rangle \le 2 - \frac{1}{40T_{\mathrm{in}}^2} \ \ (s \ne t),$$

and a softmax attention with query $a_T\varphi(s)$, $a_T = 40T_{\mathrm{in}}^2 \log(8T_{\mathrm{in}})$, and keys $\varphi(t)$ puts weight at least $7/8$ on position s.

Proof. Let k, l be the nearest integers to $r'/(2\pi)$ and $r'/(2\pi\sqrt{10})$, and put $u = r' - 2\pi k$, $v = r'/\sqrt{10} - 2\pi l$, so that $|u|, |v| \le \pi$ and $|k|, |l| \le r'$. If $(k, l) \ne (0, 0)$, then $k^2 - 10l^2$ is a nonzero integer, hence

$$\Big|k - \sqrt{10}\, l\Big| \ge \frac{1}{|k| + \sqrt{10}\,|l|} \ge \frac{1}{5r'}.$$

By the Cauchy–Schwarz inequality, $u^2 + v^2 \ge 4\pi^2/(275r'^2)$, and $1 - \cos x \ge 2x^2/\pi^2$ on $[-\pi, \pi]$ gives

$$\Gamma(r') \ge \frac{2(u^2 + v^2)}{\pi^2} \ge \frac{8}{275r'^2} \ge \frac{1}{40T_{\mathrm{in}}^2}.$$

If $k = l = 0$, then $u = r' \ge 1$ and the bound is immediate. The softmax claim follows because the logit gap is at least $a_T/(40T_{\mathrm{in}}^2) = \log(8T_{\mathrm{in}})$ and there are at most T_{in} positions. □

A ReLU "plateau" $r(z) = 2(z - 1/4)_+ - 2(z - 3/4)_+$ recovers a one-hot vector exactly from a vector whose correct coordinate is $\ge 7/8$ and whose other coordinates are $\le 1/8$. We also use the exact ReLU gate

$$G(z, c) = (z + C_{\mathrm{g}}(c - 1))_+ - (-z + C_{\mathrm{g}}(c - 1))_+ = cz \qquad (c \in \{0, 1\},\ |z| \le C_{\mathrm{g}}), \tag{36}$$

and the identity $h = (h)_+ - (-h)_+$ to copy the coordinates that an FFN must preserve (the FFNs have no residual connection, so every FFN reproduces the part of the state that later blocks need; this at most doubles the width).

The complexity of the critic is measured by the covering entropy of the whole class of Section 3.1, all parameters being free; no sub-class adapted to the unknown reward is used.

Lemma 14 (Covering number of the full class)**.** *Let $\mathcal{T} = \mathcal{T}(N_{\mathrm{blk}}, D_{\mathrm{emb}}, L_{\mathrm{FFN}}, W_{\mathrm{FFN}}, S_{\mathrm{FFN}}, B_{\mathrm{par}})$ with $B_{\mathrm{par}} \ge 1$ and $W_{\mathrm{FFN}} \ge D_{\mathrm{emb}}$, let P be its number of parameters, and let $A_0 = C(D_{\mathrm{emb}} + W_{\mathrm{FFN}} + 1)(B_{\mathrm{par}} + 1)$. For $0 < \varepsilon \le 1$, the ε-covering number of $\mathcal{T}$ in the sup norm over all valid inputs of length T_{in} satisfies*

$$\log \mathcal{N}_\infty(\varepsilon, \mathcal{T}) \le N_{\mathrm{blk}} S_{\mathrm{FFN}} \log\big(2L_{\mathrm{FFN}}(W_{\mathrm{FFN}} + 1)^2\big) + P \log\left(1 + \frac{2B_{\mathrm{par}} A_0^{CN_{\mathrm{blk}}^2(L_{\mathrm{FFN}}+2)}}{\varepsilon}\right). \tag{37}$$

Proof. There are at most $(2L_{\mathrm{FFN}}(W_{\mathrm{FFN}}+1)^2)^{S_{\mathrm{FFN}}}$ sparsity patterns per FFN. For a fixed pattern, quantize all parameters to a grid of step ϱ, which gives $(1+2B_{\mathrm{par}}/\varrho)^P$ networks. Since one-hot inputs and the positional encoding are bounded, the embedded states are bounded by $2B_{\mathrm{par}}$. If the states before a block are bounded by R_0, they are bounded by $(1+D_{\mathrm{emb}}B_{\mathrm{par}})R_0$ after the attention (its output is a convex combination of values) and by $A_0^{C(L_{\mathrm{FFN}}+2)}(R_0+1)$ after the FFN, so all states are bounded by $A_0^{CN_{\mathrm{blk}}(L_{\mathrm{FFN}}+2)}$. The softmax is 2-Lipschitz from the sup norm of the logits to the ℓ_1 norm of the weights, so the output of one block is $A_0^{C(L_{\mathrm{FFN}}+2)}(R_0+1)^C$-Lipschitz in the parameters and in the input; composing the blocks, the output is $A_0^{CN_{\mathrm{blk}}^2(L_{\mathrm{FFN}}+2)}$-Lipschitz in the parameters, and the output clipping is 1-Lipschitz. The input length T_{in} does not enter, because the attention weights sum to one. Choosing $\varrho=\varepsilon A_0^{-CN_{\mathrm{blk}}^2(L_{\mathrm{FFN}}+2)}$ gives (37). □

This is the argument of Takakura & Suzuki (2023, Theorem 5.3) and Schmidt-Hieber (2020, Lemma 5), adapted to causal attention, a learned token embedding of size $O(M)$ and the fixed positional encoding.

The batch is collected by the current actor, so the inputs are i.i.d. only given the history and the noise is only conditionally sub-Gaussian; the oracle inequality is stated in that form.

Lemma 15 (ERM oracle inequality under conditionally sub-Gaussian noise)**.** *Fix the history at the beginning of a batch. Let $U_1,\dots,U_{n_b}$ be i.i.d. inputs with law Π and $Z_t=Q(U_t)+\xi_t$ with $|Q|\le 1$, where ξ_t is conditionally mean-zero and σ-sub-Gaussian given the past and U_t. Let $\mathcal{F}$ be a class of functions with values in $[-1,1]$ whose sup-norm covering entropy is $V(\varepsilon)$, and let $\hat f$ be any empirical risk minimizer of the squared loss over $\mathcal{F}$. Then, for $n_b\ge 2$ and $\delta\in(0,1)$, with probability at least $1-\delta$,*

$$\left\|\hat f-Q\right\|_{L^2(\Pi)}^2 \le C\inf_{f\in\mathcal{F}}\|f-Q\|_{L^2(\Pi)}^2 + C(1+\sigma^2)\frac{V(n_b^{-2})+\log(2n_b/\delta)}{n_b}, \tag{38}$$

and the same bound holds in expectation with $\log(2n_b)$ in place of $\log(2n_b/\delta)$.

Proof. Let $f_0\in\mathcal{F}$ be fixed, and write $g=f-f_0$, $b_0=f_0-Q$ and Π_{n_b} for the empirical measure. The basic inequality of ERM reads

$$\Pi_{n_b}(\hat f-Q)^2 \le \Pi_{n_b}(f_0-Q)^2 + 2\Pi_{n_b}\Big[\xi(\hat f-f_0)\Big]. \tag{39}$$

The proof proceeds in three steps, starting from the basic inequality for the empirical risk minimizer. That inequality for the empirical risk minimizer leaves a noise term, an empirical squared error and an approximation term. We bound the noise term uniformly over a covering of the class by a sub-Gaussian maximal inequality, transfer the empirical squared error to its population version by Bernstein's inequality on the same covering, and combine the two with the approximation term.

Noise term. For a fixed g, the process

$$\exp\left(u'\sum_{t=1}^{n_b}\xi_t g(U_t) - \frac{\sigma^2u'^2}{2}\sum_{t=1}^{n_b}g(U_t)^2\right)$$

is a supermartingale with expectation at most one, because $g(U_t)$ is determined before ξ_t is revealed and the conditional moment generating function of ξ_t is bounded; the noise is not treated as independent of the inputs. Applying this to an n_b^{-2}-net of $\mathcal{F}$ with a union bound, and controlling the discretization error by $\max_t|\xi_t|\le C\sigma\sqrt{\log(Cn_b/\delta)}$, we obtain simultaneously for all $f\in\mathcal{F}$, with probability $1-\delta/2$,

$$2\Pi_{n_b}[\xi g] \le \frac{1}{16}\Pi_{n_b}g^2 + C\sigma^2\frac{V(n_b^{-2})+\log(C/\delta)}{n_b}. \tag{40}$$

Empirical to population. Bernstein's inequality (Boucheron et al., 2013, Theorem 2.10) for the i.i.d. inputs on the same net, with $g^2\le 4$ and $\mathbb{E}g^4\le 4\mathbb{E}g^2$, gives simultaneously

$$\Pi g^2 \le 2\Pi_{n_b}g^2 + C\frac{V(n_b^{-2})+\log(C/\delta)}{n_b}, \qquad \Pi_{n_b}b_0^2 \le 2\Pi b_0^2 + C\frac{\log(C/\delta)}{n_b}. \tag{41}$$

Conclusion. Combining (39)–(41) with $2b_0g\ge -\frac{1}{2}g^2-2b_0^2$, absorbing the empirical squared term into the left-hand side and taking the infimum over f_0 yields (38); the expectation bound follows by integrating the tail, since the loss is bounded by 4. □

Windows. Fix a step h and a lookahead $S \geq 1$, and let $J \geq 0$ be a window length. Put $s_h = \max\{0, h - J\}$ and define the *window* of step h as the set of inputs

$$\mathcal{W}_h = \big\{(m, D) : \; s_h \leq \mathrm{depth}(D) \leq h + S, \; D \subset P^*_{s_h}(m)\big\}, \tag{42}$$

i.e., the prefixes of depth between s_h and $h + S$ that agree with the correct path of their prompt up to depth s_h. The window is common to all prompts and is fixed by the step index; it depends on the unknown cells, and is a device of the proof. On the window, the cube average is a sum of few continuous functions of local coordinates.

Lemma 16 (Cube averages on a window are sums of few Hölder functions)**.** *Let $K_w \geq C_d(J + S + 2)$ with $C_d \geq 1 + 1/\ell_\star$. For $(m, D) \in \mathcal{W}_h$ of depth q, write $s = s_h$, $F^{\mathrm{old}}_{m,s} = c_1 h_1(\cdot; m) + \sum_{i \geq 2:\, L_{i-1} < s} c_i h_i(\cdot; m)$, and let $\mathcal{I}_h(m) = \{i \geq 2 : \; s \leq L_{i-1} < h + S\}$, a set of at most K_w levels. Then*

$$Q(m, D) = \frac{1}{\mu(D)} \int_D F^{\mathrm{old}}_{m,s} \, \mathrm{d}\mu + \sum_{i \in \mathcal{I}_h(m):\; L_{i-1} < \mathsf{q},\; D \subset C_{i-1}(m)} \frac{c_i}{\mu(D)} \int_D h_i(\cdot; m) \, \mathrm{d}\mu \, . \tag{43}$$

Moreover, the components of (43) *are smooth at the following scales.*

(a) (Old component.) *In the local coordinates of $P^*_s(m) = b_s + 2^{-s}\Omega$, the map $z \mapsto F^{\mathrm{old}}_{m,s}(b_s + 2^{-s} z)$ is α-Hölder on $[0, 1]^d$ with $|\cdot|_\alpha$ at most $C a_s$ and C^α norm at most $C a_s$ after subtracting its value at zero, and $a_s \leq C K_w^\gamma a_{h+S}$.*

(b) (Window components.) *In the local coordinates of $C_{i-1}(m)$, the primitive $h_i(\cdot; m)$ is α-Hölder with $|\cdot|_\alpha$ at most λ, and $c_i \leq C a_s$ for every $i \in \mathcal{I}_h(m)$.*

(c) (Averaging preserves smoothness.) *Fix a depth q and a cube A of depth $p \leq \mathsf{q}$ in whose local coordinates a function g is α-Hölder with seminorm $|g|_\alpha$. Then the cube average*

$$z \mapsto \int_{[0,1]^d} g\Big(z + 2^{-(\mathsf{q} - p)} u\Big) \, \mathrm{d}u$$

is α-Hölder in the lower-left corner z with seminorm at most $|g|_\alpha$. The number of depths q occurring in the window is at most $J + S + 1 \leq K_w$.

Proof. A level $i \geq 2$ with $L_{i-1} \geq \mathsf{q}$ has support $C_{i-1}(m)$ of depth at least q, which is either inside D or disjoint from D; in both cases its contribution to $Q(m, D)$ vanishes by **(A4)** and Lemma 7. A level with $s \leq L_{i-1} < \mathsf{q}$ has support of depth $L_{i-1} < \mathsf{q}$, which either contains D or is disjoint from it; only the former levels contribute, which gives (43). The number of levels with $L_{i-1} \in [s, h + S)$ is at most $(h + S - s)/\ell_\star + 1 \leq (J + S)/\ell_\star + 1 \leq K_w$ by **(A3)**. For (a), the derivative-by-derivative scaling of Lemma 10, **(A1)** and the geometric series in Lemma 9 give a C^α norm after subtraction of a constant at most $C\lambda\Big(c_1 2^{-\beta s} + \sum_{i \geq 2:\, L_{i-1} < s} c_i 2^{-\beta(s - L_{i-1})}\Big) \leq C c_{k(s)} \leq C a_s$ (the first level is included with $L_0 = 0$), and $a_s / a_{h+S} \leq C((h + S + 1)/(s + 1))^\gamma \leq C(J + S + 1)^\gamma \leq C K_w^\gamma$ by (14). (b) is **(A1)** together with $c_i \leq c_{k(s)} \leq C a_s$ for $L_{i-1} \geq s$. For (c), differentiation under the integral gives $D^s \int g(z + vu) \mathrm{d}u = \int D^s g(z + vu) \mathrm{d}u$ for $|s| \leq k$; averaging contracts each derivative sup norm and each top-order τ-Hölder seminorm. The same holds for the function-value β-Hölder seminorm, so subtraction of a constant again bounds the full local C^α norm. □

The decomposition uses one coordinate system per level: approximating all levels in the coordinates of a single anchor cube would produce Hölder seminorms of order $2^{\alpha J}$.

Lemma 17 (Window class: approximation and covering)**.** *For integers $K_w, N, T_{\mathrm{in}} \geq 1$, let $\mathcal{T}_{\mathrm{win}}(K_w, N, T_{\mathrm{in}})$ be the class $\mathcal{T}(N_{\mathrm{blk}}, D_{\mathrm{emb}}, L_{\mathrm{FFN}}, W_{\mathrm{FFN}}, S_{\mathrm{FFN}}, B_{\mathrm{par}})$ of Section 3.1 with input length T_{in} and the following architecture parameters. Then the class has the stated capacity, covering entropy and approximation power:*

(a) (Architecture.) *It suffices to take*

$$N_{\mathrm{blk}} \leq C K_w, \quad D_{\mathrm{emb}} \leq C K_w, \quad L_{\mathrm{FFN}} \leq C \ell_N, \quad W_{\mathrm{FFN}} \leq C\big[M K_w^2 (N + 1) + K_w^3\big], \tag{44}$$

$$S_{\mathrm{FFN}} \leq C\big[M K_w^2 (N + 1) \ell_N + K_w^3\big], \qquad B_{\mathrm{par}} \leq C\big[M + K_w + T_{\mathrm{in}}^2 \log(8 T_{\mathrm{in}}) + N + 1\big].$$

(b) (Covering entropy.) *Its covering entropy satisfies*

$$V_N(u) := \log \mathcal{N}_\infty(u, \mathcal{T}_{\mathrm{win}}(K_w, N, T_{\mathrm{in}})) \le CMK_w^6(N+1)\ell_N^2 \log \frac{CMT_{\mathrm{in}}K_w(N+1)}{u}. \tag{45}$$

(c) (Approximation on the window.) *For every step h with $T_{\mathrm{in}} \ge d(h+S)+3$ and $K_w \ge C_d(J+S+2)$, there is a comparator $F_{0,h} \in \mathcal{T}_{\mathrm{win}}(K_w, N, T_{\mathrm{in}})$ (depending on the unknown reward) with*

$$\sup_{(m,D)\in\mathcal{W}_h} |F_{0,h}(m,D) - Q(m,D)| \le CK_w^{\gamma+1} a_{h+S} N^{-\alpha/d}, \tag{46}$$

(d) (L^2 form.) *Consequently, for every law Π of inputs (m, D) with $\Pi(\mathcal{W}_h^c) =: \Lambda$,*

$$\|F_{0,h} - Q\|^2_{L^2(\Pi)} \le CK_w^{2\gamma+2} a_{h+S}^2 N^{-2\alpha/d} + 4\Lambda. \tag{47}$$

Proof. The proof follows the chain announced at the beginning of this subsection, in three steps: first we approximate each of the finitely many Hölder components produced by Lemma 16 by a ReLU network of size N; then we realize the sum of these networks inside the Transformer class, using the fixed positional encoding and Lemma 13 to route each record to the component it needs; finally we read off the architecture parameters and apply Lemma 14 to bound the covering entropy.

Approximation of the components. By Lemma 16, on the window $Q(m,\cdot)$ is a sum of at most K_w+1 components, the old sum and those levels of $\mathcal{I}_h(m)$ whose support contains the query cube, and for each of the at most K_w query depths q every component is an α-Hölder function of the d local coordinates of the query cube inside its support cube, with $|\cdot|_\alpha$ at most Ca_s and C^α norm at most Ca_s after subtracting its value at zero; the side length $2^{-(\mathsf{q}-p)}$ is a discrete branch, not a continuous variable. Subtracting the value at the origin and dividing by Ca_s sends each component into the unit α-Hölder ball of $[0,1]^d$, and by Schmidt-Hieber (2020, Theorem 5) every function of that ball is approximated in sup norm to accuracy $CN^{-\alpha/d}$ by a network of

$$\Psi(C\ell_N,\ CN,\ CN\ell_N,\ C)$$

(the internal precision integer of that theorem is set to $C\log(N+1)$, with constants depending on α, d; the theorem covers every $\alpha > 0$, including the integer $C^{k,1}$ cases). Per prompt there are $O(K_w^2)$ such functions, one for each component and query depth, of which at most K_w+1 are active on a given input, so the sum of the rescaled approximations, clipped to $[-1,1]$, satisfies (46) by (a)–(c) of Lemma 16. The approximation holds on the whole continuous range of the local coordinates; no interpolation of observed query values is involved.

Realization with the fixed positional encoding. Let φ be the four coordinates (34) of the positional encoding and $a_T = 40T_{\mathrm{in}}^2 \log(8T_{\mathrm{in}})$. The comparator is assembled from the following blocks.

(1) *Token embedding.* The embedding of each token contains its category (prompt, 0, 1, `STOP`, `PAD`, `OUT`) as constants spaced by at least 4 in coordinates separate from φ. Each positional shift has magnitude at most 1, so the category intervals are disjoint and the first FFN recovers the category exactly by plateaus; it preserves φ and initializes unused working coordinates to zero. By the gate (36) with the prompt category, the same FFN zeroes at all positions other than 1 the *prompt information*, namely the index m, the $d(h+S-s_h)$ bits of the correct path of prompt m at the depths $s_h+1,\dots,h+S$, and the level-boundary flags $\mathbb{1}_{L_{i-1}=j}$ for the depths j of the window and $i \in \mathcal{I}_h(m)$, that is, $O(K_w)$ numbers; the positional encoding of position 1 is subtracted in this embedding. These unknown bits and flags are parameters of the comparator, not information available to the learner.

(2) *Broadcasting.* The second block has zero attention logits, i.e. uniform causal attention, and a value map that copies the prompt information, multiplied by T_{in}, to fresh coordinates. Since only position 1 carries it, the readout position T_{in} receives exactly the stored values; other positions receive multiples T_{in}/t, which are never read again.

(3) *Reading the query cube.* Then K_w blocks read the tokens at the absolute positions

$$ds_h + 2,\ \dots,\ d(h+S)+2,$$

one block per position: the query of the readout position is $a_T\varphi(t)$ for the target position t, the keys are $\varphi(t')$ and the values are the token categories. By Lemma 13 the target receives weight at least $7/8$, and a plateau recovers its category exactly; the result is stored in a fresh coordinate and the temporary coordinates are reset, so that earlier reads do not contaminate later keys and values. Since the FFNs have no residual connection, each of them copies the coordinates that later blocks need, using $x = (x)_+ - (-x)_+$; this at most doubles the width.

(4) *Computing the arguments.* From the read tokens, an FFN of size $O(K_w^3)$ computes the query depth q from the position of STOP; the support gates $\mathbb{1}_{D\subset C_{i-1}(m)}$ for $i \in \mathcal{I}_h(m)$, as ReLU functions of the number of mismatches between the read bits and the stored path bits up to depth L_{i-1}; and the local coordinates

$$z_j = \sum_{r'=p+1}^{\mathsf{q}} b_{r',j} 2^{-(r'-p)}, \qquad j \in [d],$$

of the query cube in the coordinates of each candidate support cube of depth p, namely $p = s_h$ for the old sum and $p = L_{i-1}$ for the level i.

(5) *Evaluating and selecting.* A shared FFN of width $O(MK_w^2(N+1) + K_w^3)$ evaluates in parallel the $O(MK_w^2)$ approximating networks of all prompts, components and depths, and the exact gate (36) with the product of the indicators $\mathbb{1}_{m'=m}$, $\mathbb{1}_{\mathsf{q}'=\mathsf{q}}$ and the support gate selects the active ones and adds them, so that the approximation error is not multiplied by M or K_w; the output is clipped to $[-1, 1]$.

All information flows to the readout position, in accordance with the causal mask. On inputs outside the window, that is, a prefix that disagrees with the correct path before depth s_h or a depth below s_h, the comparator outputs some value in $[-1, 1]$; no detection of such inputs is claimed, and their error, at most 2 in absolute value, is charged to Λ in (47). The construction fits (44): the number of blocks is $O(K_w)$, the state dimension $O(K_w)$, the FFN depth $O(\ell_N)$, and the weights are bounded by $O(M + K_w + T_{\text{in}}^2 \log(8T_{\text{in}}) + N + 1)$.

Covering. Counting the dense token embedding, the attention matrices, the readout and the nonzero FFN weights, the number of parameters is at most $D_{\text{emb}}(M+5) + 3N_{\text{blk}}D_{\text{emb}}^2 + N_{\text{blk}}S_{\text{FFN}} \le CMK_w^4(N+1)\ell_N$, and the sparsity patterns are part of the class; Lemma 14 with these capacities gives (45) (the input length enters only through the logarithm of the weight bound). (47) follows from (46) and $|F_{0,h} - Q| \le 2$. □

The learner is not restricted to the comparator: all internal weights of the class $\mathcal{T}_{\text{win}}$ are fitted by ERM, and the comparator is used only to bound the approximation error in Lemma 15.

D Proof of the Large-Budget Bound of Theorem 3

Proof strategy. The algorithm of the large-budget bound of Theorem 3 consists of two stages, and the proof follows that division. The *front phase* is the sequence of exploration iterations $h < H$: it refines the score along the correct path and is analyzed in Appendix D.2, whose main result is Proposition 18, a high-probability guarantee that every update stays birth-accurate together with a bound on the total number of queries that the phase consumes. The *final refinement step* is the single iteration $h = H$: conditionally on the success event of the front phase, it keeps the score W_H fixed, samples from the Gibbs policy at the target temperature ρ, fits one more critic at a variable query depth and outputs the clipped score. It is analyzed in Appendix D.3, whose main result is Proposition 24, a conditional regret bound that still exhibits the bias–variance tradeoff in the refinement resolution. Appendix D.4 then assembles the two propositions: it balances the resolution, uses the remaining budget $b \ge n/2$, substitutes the bound on H and adds the contribution of the front-phase failure event. Schematically,

Proposition 18 (front phase) $\implies$ Proposition 24 (final refinement) $\implies$ Theorem 3.

D.1 Hyperparameters of the Large-Budget Setting

This section concerns the large-budget setting of Section 3.2: $\ell_{\text{look}} = S := R + 1$, $H_{\text{tot}} = H + 1$, and the last iteration $h = H$ is the refinement step (6) with $I_H \equiv 1$ and $\vartheta_H = \vartheta_{H+1} = \rho$; the small-

regularization setting is treated in Appendix E. We call the exploration iterations $h < H$ the *front phase* and the final iteration the *final refinement step*, and use the following names: the batch size, class, actor and critic of the final refinement step are $b := b_H$, $\mathcal{T}_{\mathrm{ref}} := \mathcal{T}_H$, $P_{\mathrm{start}} := P_H = P_{W_H,\rho}$ and $\widehat{T}_{\mathrm{ref}} := \widehat{T}_H$; the query cube of the final refinement step is $D_{\mathrm{ref}}(m,y) := \mathsf{Query}_H(m,y) = D_{q(m,y)+S+r}(y)$; and the clipped score is $v := W_{H+1}$, so that $\widehat{P} = P_{v,\rho}$. In the front phase, the query map is written $\Psi_h := \mathsf{Query}_h$ of (5), and $\Delta_h = 2$, so that line 6 of Algorithm 1 reads $W_{h+1}(m,y) = \widehat{T}_h(m, D_{h+S}(y))$ on the selected cubes and $W_{h+1} = W_h$ elsewhere.

The quantities listed in Table 3 are specific to this proof; the schedules that both upper bounds share are those of Table 2, whose values are fixed below for the large-budget setting.

Table 3: Quantities used only in the proof of the large-budget bound of Theorem 3.

Symbol	Meaning
H	number of front-phase iterations
R, $S = R+1$	refinement depth of the lookahead; the lookahead itself
η, r_{mom}	birth-accuracy level; moment order used in the regret transfer
δ_F, δ_h	total and per-phase failure budgets
T_{in}	input length of the front-phase critic
b, r, N_{ref}	batch, query depth and approximation size of the final refinement step
$\mathcal{L}$	logarithmic factor $\log\bigl(e+n+M+\rho^{-1}\bigr)$

Given (n, M, ρ), the algorithm fixes the following quantities before the first query. The structural constants $C_H, C'_H, C_{\mathrm{mom}}, c_\eta, t_0, C_J, C_N, C_b, D_N, D_b$ and C_{cal} (the constant of (62)) are chosen in the proofs below, in the order indicated there; in particular $c_\eta \le \eta_0$ with η_0 the constant of Lemma 12.

$$
\begin{aligned}
H &= \lceil C_H K + C'_H \mathcal{L} \rceil\,, & r_{\mathrm{mom}} &= \lceil C_{\mathrm{mom}} \mathcal{L} \rceil,\\
\eta &= \frac{c_\eta}{r_{\mathrm{mom}}}, & \delta_F &= \frac{\rho^2}{(n+1)^2}, \qquad \delta_h = \frac{\delta_F}{H},\\
R &= \left\lceil \frac{1}{\beta} \log_2 \frac{2C_{\mathrm{cal}}}{\eta} \right\rceil, \quad S = R+1, & \varepsilon_h &= \frac{\eta}{4}\, a_{h+S} \quad (0 \le h < H),\\
\vartheta_h &= t_0\, c_{1+\lfloor h/\ell_\star \rfloor} \quad (0 \le h \le H-1), & \kappa_0 &= \frac{p_*}{2},
\end{aligned}
\tag{48}
$$

where C_H, C'_H are large enough that $H \ge \bar{\ell}\lceil K \rceil + 1$ and $S \le H$ (recall $S = O(\log \mathcal{L})$), and p_* is the constant of Lemma 12. The accuracy ε_h required of the critic of step h is tied to the depth $h+S$ of the cubes that this step queries and writes into the score (*query-depth accuracy*); no common absolute accuracy is required of all steps. For the front phase, the window length, the window size, the input length, the approximation size and the batch sizes are

$$
\begin{aligned}
J &= \lceil C_J \mathcal{L} \rceil, & K_w &= C_d(J+S+2), \\
T_{\mathrm{in}} &= d(H+S)+3, & N &= \left\lceil C_N M^{\frac{d}{2\alpha}} \mathcal{L}^{D_N} \right\rceil,\\
b_h &= \left\lceil C_b(1+\sigma^2)\, M^{\zeta_M}\, a_{h+S}^{-2}\, \mathcal{L}^{D_b} \right\rceil \quad (0 \le h < H),
\end{aligned}
\tag{49}
$$

and $\mathcal{T}_h$ is the class $\mathcal{T}_{\mathrm{win}}(K_w, N, T_{\mathrm{in}})$ of Lemma 17 (whose input-length argument is written T_{in} there), the same for all $h < H$; the window parameter J enters only the analysis (through the comparator of Lemma 17), not the algorithm. For the final refinement step, the batch size is $b = n - \sum_{h<H} b_h$, and the integer approximation size $N_{\mathrm{ref}} \ge 1$ is chosen with

$$
N_{\mathrm{ref}} \asymp \max\left\{1, \left(\frac{b\rho^2}{(1+\sigma^2)M(H+1)\mathcal{L}^{D_{\mathrm{E}}}}\right)^{\frac{d}{2\alpha+d}}\right\},
\tag{50}
$$

the query depth increment is $r = \lceil \alpha \log_2 N_{\mathrm{ref}}/(\beta d) \rceil$, the input length is $T_{\mathrm{ref}} = d(H+S+r)+3$, and $\mathcal{T}_{\mathrm{ref}}$ is the class of Lemma 27 with these H, S, N_{ref}. The constants $C_{\mathrm{F}}, D_{\mathrm{F}}$ of the large-budget bound of Theorem 3 are such that $\sum_{h<H} b_h \le n/2$ whenever $n \ge C_{\mathrm{F}}(1+\sigma^2) M^{\zeta_M} \rho^{-(2+1/\gamma)} \mathcal{L}^{D_{\mathrm{F}}}$

(Proposition 18), so that $b \geq n/2$. The second condition of Theorem 3 is this requirement written in terms of ρ, with $c_{\log} = \gamma D_{\mathrm{F}}/(2\gamma+1)$ and implicit constant $[C_{\mathrm{F}}(1+\sigma^2)]^{\gamma/(2\gamma+1)}$. If the budget is below this threshold, the algorithm outputs the uniform policy; this case is outside the large-budget bound of Theorem 3.

D.2 The Front Phase

The goal of this subsection is the following guarantee for the exploration iterations $h < H$. It is stated here, before the lemmas that prove it, because all of them are ingredients of this one statement.

Proposition 18 (Front-phase guarantee and cost)**.** *With the parameters of Appendix D.1,*

$$\Pr(\mathcal{E}) \geq 1 - \delta_F,$$
$$\sum_{h<H} b_h \leq C(1+\sigma^2) M^{\zeta_M} H^{2\gamma+1} \mathcal{L}^{D_b} \leq \frac{C_{\mathrm{F}}}{2}(1+\sigma^2) M^{\zeta_M} \rho^{-(2+1/\gamma)} \mathcal{L}^{D_{\mathrm{F}}}, \tag{51}$$

where $\mathcal{E}$ is the event (52) *that every critic of the front phase is accurate on its update inputs.*

The proof is an induction over the phases, and the lemmas below are its steps. It is not a single chain: after the bookkeeping of Lemma 19 the argument branches into Lemma 20 and Lemma 21, and the two branches rejoin at Lemma 23:

$$\mathcal{E}_h \longrightarrow \begin{cases} \text{Lemma 20: true-path progress} \\ \quad \to \text{ Lemma 22: localization} \\ \quad \to \text{ Lemmas 15 and 17: joint risk,} \\ \\ \text{Lemma 21: fine-query mass } p_S, \end{cases} \longrightarrow \text{Lemma 23} \longrightarrow \mathcal{E}_{h+1}.$$

Proposition 18 iterates this implication over the H phases, sums the conditional failure probabilities and adds up the batch sizes.

The lemmas of this subsection are stated for a general lookahead $S \geq 1$, a general accuracy sequence $\varepsilon_h = \frac{\eta}{4} a_{h+S}$ and general batch sizes b_h; they are used with the parameters of Appendix D.1 here and with $S = 1$ in Appendix E. We analyze the scores as functions on the *analysis partition* generated by the updates: initially $\mathcal{D}_0(m) = \{\Omega\}$, and at step h the selected cubes are replaced by their depth-$(h+S)$ sub-cubes. The algorithm never forms this partition.

Lemma 19 (Analysis partition with lookahead)**.** *For every history and every h, the following hold.*

- *(i) $W_h(m,\cdot)$ is constant on the cubes of a partition $\mathcal{D}_h(m)$ of Ω into dyadic cubes.*
- *(ii) Every cube $D \in \mathcal{D}_h(m)$ except the initial root was created at a step $r(D) < h$ and has depth $r(D)+S$.*
- *(iii) The set $\{y : P_h(D_h(y) \mid m) \geq \kappa_0\}$ of responses is a union of cubes of $\mathcal{D}_h(m)$ and of depth-h cubes, of which there are at most $1/\kappa_0$; these are the* selected *cubes.*
- *(iv) $\mathcal{D}_{h+1}(m)$ is obtained from $\mathcal{D}_h(m)$ by splitting the selected cubes into their depth-$(h+S)$ sub-cubes, the* fine queries *of step h; the other cubes are left untouched and keep their score.*

In particular the partitions are refined and never coarsened, so the birth step $r(D)$ and the birth depth $r(D)+S$ of every current cube are well defined.

Proof. If a cube $D' \in \mathcal{D}_h$ has depth at most h, the actor is uniform on D', so all depth-h cubes inside D' have the same mass and are selected together; if D' is deeper than h, it lies inside one depth-h cube. Hence the mask I_h is constant on the cubes of $\mathcal{D}_h$. The update of line 6 with $\Delta_h = 2$ replaces the score on the selected cubes by the values $\widehat{T}_h(m,D)$ on their depth-$(h+S)$ sub-cubes D, which have depth $h+S > h-1+S \geq \max_{D' \in \mathcal{D}_h} \mathrm{depth}(D')$, and keeps it elsewhere. The count follows since the masses of disjoint cubes sum to at most one. □

Let $\mathcal{E}_h$ be the event that all critics of the steps $h' < h$ are accurate on their update inputs:

$$\mathcal{E}_h = \left\{ \left| \widehat{T}_{h'}(m,D) - Q(m,D) \right| \leq \varepsilon_{h'} \text{ for all } h' < h,\ m \in [M], \text{ and fine queries } D \text{ of step } h' \right\}, \tag{52}$$

and $\mathcal{E} = \mathcal{E}_H$; $\mathcal{E}_h$ is determined by the history at the beginning of step h. On $\mathcal{E}_h$, every cube $D \in \mathcal{D}_h(m)$ created at step $r(D)$ satisfies $|W_h(m, D) - Q(m, D)| \le \varepsilon_{r(D)} = \frac{\eta}{4} a_{\mathrm{depth}(D)}$, i.e., $W_h(m, \cdot)$ is η-birth-accurate on $\mathcal{D}_h(m)$ in the sense of Definition 11 (the root is updated at step 0, so no cube keeps the initial value 0 for $h \ge 1$).

If the scores written so far are accurate, the correct cube carries enough actor mass to be refined, and the mass of the responses that left the path much earlier is exponentially small.

Lemma 20 (The correct cube is refined at every step)**.** *Let $\eta \le \eta_0$. On $\mathcal{E}_h$, for every m and $h \ge 1$, the correct cube $P_h^*(m)$ is a union of cubes of $\mathcal{D}_h(m)$, all created at step $h - 1$,*

$$P_h(P_h^*(m) \mid m) \ge p_*, \qquad P_h\big(\big(P_{h-j}^*(m)\big)^c \;\big|\; m\big) \le Ce^{-cj} \quad (0 \le j \le h), \tag{53}$$

and, with $\kappa_0 = p_*/2$, the correct cube $P_h^*(m)$ and all its ancestors are selected at step h. For $h = 0$ the root is selected.

Proof. By induction: if the correct parent $P_{h-1}^*(m)$ was selected at step $h - 1$, it was split into depth-$(h - 1 + S)$ cubes, so $P_h^*(m)$ (of depth $h \le h - 1 + S$) is a union of these cubes. On $\mathcal{E}_h$ the score is η-birth-accurate on $\mathcal{D}_h(m)$, and the temperature satisfies (29) by (33), so Lemma 12 gives (53). Every ancestor of $P_h^*(m)$ has mass at least $p_* \ge \kappa_0$, hence passes the mass check. □

The previous lemma says which cube is refined; the next says how much probability each cube created by that refinement receives.

Lemma 21 (Every fine query has a positive probability)**.** *On $\mathcal{E}_h$, for every m, every selected depth-h cube A' and every depth-$(h + S)$ cube $D \subset A'$,*

$$P_h(D \mid m) \ge p_S := \kappa_0 e^{-C_S S} 2^{-dS}, \tag{54}$$

and every response that enters D is queried at D. In particular $p_S^{-1} \le C\mathcal{L}^{D_p}$ for a fixed D_p when $S = O(1 + \log \mathcal{L})$, and $p_S = \kappa_0 2^{-d}$ when $S = 1$.

Proof. If A' is contained in one cube of $\mathcal{D}_h(m)$, the actor is uniform on A' and $P_h(D \mid m) = 2^{-dS} P_h(A' \mid m) \ge \kappa_0 2^{-dS}$. Otherwise the cubes of $\mathcal{D}_h(m)$ inside A' have depths between h and $h + S - 1$, their cube averages differ by at most $CSc_{k(h)}$ by Lemma 9 (with $R + 1 = S$), and their scores are within $\frac{\eta}{4} a_h \le C\eta c_{k(h)}$ of the averages by birth accuracy; as $\vartheta_h \asymp c_{k(h)}$, the density of P_h on A' is within a factor $e^{C_S S}$ of uniform, which gives (54). A response entering D passes the mass checks of all ancestors of D up to depth h (their masses are at least that of A'), so it is not stopped and is queried at depth $h + S$, i.e., at D. The bound holds for every selected cube, not only for the correct one. □

Up to an exponentially small leakage, the queried inputs of step h fall in the window $\mathcal{W}_h$, which is the finite region on which Lemmas 16 and 17 operate.

Lemma 22 (Window localization)**.** *Let Π_h be the joint law of the queried input $(X_t, \Psi_h(X_t, Y_t))$ of step h (a fresh prompt and one on-policy response). On $\mathcal{E}_h$,*

$$\Lambda_h := \Pi_h(\mathcal{W}_h^c) \le Ce^{-cJ}, \tag{55}$$

and $\Lambda_h = 0$ if $s_h = 0$.

Proof. Every query has depth at most $h + S$. If the response lies in $P_{s_h}^*(m)$, all its prefixes up to depth s_h are ancestors of $P_h^*(m)$, which pass the mass check by Lemma 20, so the response is not stopped before depth s_h and its queried cube, of depth at least s_h, is contained in $P_{s_h}^*(m)$, i.e., the input lies in $\mathcal{W}_h$. Hence $\mathcal{W}_h^c \subset \{Y \notin P_{s_h}^*(X)\}$, whose probability is at most $Ce^{-c(h-s_h)} \le Ce^{-cJ}$ by (53). □

The oracle inequality bounds the risk averaged over the query law, whereas the induction needs accuracy at *every* cube written into the score; each of them has probability at least p_S, so a joint risk of order $p_S \varepsilon_h^2 / M$ suffices, and this is where the factor M^{ζ_M} is paid.

Lemma 23 (Joint risk of the full-class ERM and accuracy on the update inputs)**.** *Let $\mathcal{T}_h = \mathcal{T}_{\mathrm{win}}(K_w, N, T_{\mathrm{in}})$ with $K_w \geq C_d(J+S+2)$ and $T_{\mathrm{in}} \geq d(h+S)+3$, and let $\widehat{T}_h$ be the ERM of line 5 on the b_h records of step h. Conditionally on a history in $\mathcal{E}_h$, with probability at least $1-\delta_h$,*

$$\mathsf{R}_h := \left\|\widehat{T}_h - Q\right\|^2_{L^2(\Pi_h)} \leq C\left[K_w^{2\gamma+2} a_{h+S}^2 N^{-2\alpha/d} + \Lambda_h + (1+\sigma^2)\frac{V_N(b_h^{-2}) + \log(2b_h/\delta_h)}{b_h}\right], \tag{56}$$

and, deterministically, for every m and every fine query D of step h,

$$\left|\widehat{T}_h(m,D) - Q(m,D)\right|^2 \leq \frac{M}{p_S}\mathsf{R}_h. \tag{57}$$

Proof. Conditionally on the history at the beginning of step h, the actor, the query rule and the class are fixed, the inputs $(X_t, \Psi_h(X_t, Y_t))$ are i.i.d. with law Π_h (a fresh exogenous prompt and an independent rollout), the targets are $Z_t = Q(X_t, \Psi_h(X_t, Y_t)) + \xi_t$ with conditionally σ-sub-Gaussian noise, and all records, including the stopped ones, enter the unweighted loss. Lemma 15 with $\mathcal{F} = \mathcal{T}_h$ and the comparator of Lemma 17 gives (56), using (47) with $\Lambda = \Lambda_h$ and (45). For (57): the joint probability of the input (m, D) is $\Pi_h(\{(m,D)\}) = P_h(D \mid m)/M \geq p_S/M$ by Lemma 21, and $\mathsf{R}_h \geq \Pi_h(\{(m,D)\})\left|\widehat{T}_h(m,D) - Q(m,D)\right|^2$. □

The factor M in (57) is paid in full: the joint L^2 risk is converted into the accuracy on every update input of every prompt, and this is what makes the batch sizes grow like M^{ζ_M} rather than M; no prompt-wise estimation guarantee is used.

Proof of Proposition 18. The accuracy statement is proved by induction over the phases: conditionally on a history in $\mathcal{E}_h$, the parameters are chosen so that the two error terms of the joint risk bound (56) are below the level that (57) needs in order to give the required accuracy at every node updated at phase h, which is exactly $\mathcal{E}_{h+1}$; summing the conditional failure probabilities over the phases gives the first half of (51). The cost statement then only adds up the batch sizes and uses the bound on H.

Accuracy. Condition on a history in $\mathcal{E}_h$. By (57) it suffices that the right-hand side of the joint risk bound (56) be at most

$$\mathsf{r}_h := c\,\frac{p_S \eta^2 a_{h+S}^2}{M}, \tag{58}$$

and we check that each of its three terms is at most r_h separately.

- *Leakage.* By Lemma 22 and $H \leq n$, choosing C_J large gives

$$\Lambda_h \leq Ce^{-cJ} \leq c\,\frac{p_S \eta^2 a_{H+S}^2}{M} \leq \mathsf{r}_h \qquad (h < H),$$

  where we used $\eta^{-1} = O(\mathcal{L})$, $p_S^{-1} = O(\mathcal{L}^{D_p})$, $a_{H+S}^{-2} = O(H^{2\gamma})$ and $M \leq n$ on the range of the theorem.

- *Approximation.* With $K_w = O(\mathcal{L})$ and $N = \lceil C_N M^{d/(2\alpha)} \mathcal{L}^{D_N} \rceil$ for D_N large, the approximation term $CK_w^{2\gamma+2} a_{h+S}^2 N^{-2\alpha/d}$ of (56) is at most r_h.

- *Estimation.* Since $b_h \leq n$, (45) and the choice of δ_h give

$$V_N(b_h^{-2}) \leq CM^{1+\frac{d}{2\alpha}} \mathcal{L}^{D_N+9}, \qquad \log\frac{2b_h}{\delta_h} \leq C\mathcal{L},$$

  so that with $D_b \geq D_N + D_p + 11$ and C_b large the estimation term of (56) is at most r_h as well.

Hence, by (57), with conditional probability at least $1-\delta_h$ every fine query of step h is $\varepsilon_h = \frac{\eta}{4} a_{h+S}$-accurate, i.e. $\mathcal{E}_{h+1}$ holds. Summing the conditional failure probabilities up to the first failure over the H steps,

$$\Pr(\mathcal{E}^c) \leq \sum_{h<H} \delta_h = \delta_F.$$

No independence between the noise and the future occurrence of inputs is assumed, and no conditioning on future successes is used.

Cost. Since $S \le H$ and $a_j^{-2} \asymp (j+1)^{2\gamma}$,

$$\sum_{h<H} b_h \le C(1+\sigma^2) M^{\zeta_M} \mathcal{L}^{D_b} \sum_{h<H} a_{h+S}^{-2} \le C(1+\sigma^2) M^{\zeta_M} H^{2\gamma+1} \mathcal{L}^{D_b},$$

which is the first inequality of (51), the rounding in (49) being absorbed into C; the second follows from $H \le CK\mathcal{L}$ with $C_{\mathrm{F}}, D_{\mathrm{F}}$ large. All records are counted: stopped responses, queries outside the window and the batches after a failure. □

D.3 The Final Refinement Step

Condition on the successful front-phase event $\mathcal{E}$ of Proposition 18. No query, critic fit or score update takes place between the two stages: the front-phase score W_H is kept fixed and the Gibbs temperature is set to the target regularization level, so that the responses of the last iteration are drawn from $P_{\mathrm{start}} = P_{W_H,\rho}$. The goal of this subsection is the following conditional guarantee, expressed in terms of the resolution N_{ref} chosen in (50), to display the approximation and estimation errors separately.

Proposition 24 (Final-refinement guarantee)**.** *With the parameters of Appendix D.1 on the large-budget range of Theorem 3, including the refinement resolution N_{ref} chosen in* (50)*, the output $\widehat{P} = P_{v,\rho}$ of Algorithm 1 satisfies*

$$\mathbb{E}\Big[\mathcal{R}_\rho(\widehat{P}) \;\Big|\; \mathcal{E}\Big] \le C\rho N_{\mathrm{ref}}^{-\frac{2\alpha}{d}} + \frac{C(1+\sigma^2)M(H+1)(N_{\mathrm{ref}}+1)}{b\rho}\mathcal{L}^{D_{\mathrm{E}}} + \frac{C\delta_C}{\rho}, \tag{59}$$

where b is the batch of the final iteration and δ_C is the mass of the deep core (64)*.*

The three terms are the approximation error of a resolution-N_{ref} fit, the estimation error of one critic fitted on b records, and the mass of the deep core. The proof is the chain

Lemma 25: pointwise calibration → Lemma 26: Gibbs mass and moments
→ Lemma 27: final critic estimation
→ Lemma 28: continuous reward and clipping
→ Lemma 29: regret transfer.

Calibration and Gibbs control. For the rest of the proof we fix a good front-phase history (the event $\mathcal{E}$) and a prompt m; all conditional expectations $\mathbb{E}[\,\cdot \mid \mathcal{E}]$ below are taken given that history. We write

$$P_j = P_j^*(m) \;\; (0 \le j \le H), \qquad A_j = P_j \setminus P_{j+1} \;\; (0 \le j < H),$$
$$C_{\mathrm{core}} = P_H, \qquad P_{\mathrm{start}} = P_{W_H,\rho}.$$

Only for the proof, consider the analysis partition $\mathcal{D}_H$ of the final score $W_H(m,\cdot)$. By Lemma 19, each update refines the previous partition and never coarsens it, so the last update index q is constant on each final cube $A' \in \mathcal{D}_H$, and, on $\mathcal{E}$,

$$\mathrm{depth}(A') = q(A') + S, \quad W_H(A') = \widehat{T}_{q(A')}(m, A'), \quad |W_H(A') - Q(m, A')| \le \tfrac{\eta}{4} a_{q(A')+S}. \tag{60}$$

Since the root is updated at $h = 0$, no cube keeps the initial value 0, and $W_H \le G_m + \frac{\eta}{4} a_{q+S}$ everywhere. The algorithm never enumerates $\mathcal{D}_H$; the index $q(m,y)$ of (6) equals $q(A')$ for the cube $A' \ni y$, and is evaluated for each sample from the saved actors.

A response that leaves the correct path at depth j lies in a cube last updated at some step $q \ge j$, so the error of the held score there is measured at the scale a_q of that update rather than at a common scale.

Lemma 25 (Calibration of the held score)**.** *On $\mathcal{E}$, for every final cube $A' \in \mathcal{D}_H$ contained in the shell A_j ($j < H$),*

$$q(A') \ge j, \tag{61}$$

and for every $y \notin C_{\mathrm{core}}$ with last update index $q = q(m,y)$,

$$|f(y;m) - W_H(m,y)| \le \frac{\eta}{4} a_{q+S} + C_{\mathrm{cal}} a_q 2^{-\beta S} \le \eta a_q. \tag{62}$$

Proof. By Lemma 20 the correct parent P_j is selected at step j, so the whole shell $A_j \subset P_j$ satisfies the mass condition at step j and is updated at depth $j+S$; the holding of later steps never moves a query depth back, which gives (61), and a final cube of depth $q+S \geq j+1$ is contained in one shell by dyadic consistency. Let $i(j) = \min\{i : L_i > j\}$, so that $f = f^{[i(j)]}$ on A_j. By (27) (the cube A' has side $2^{-(q+S-j)}$ in the local coordinates of P_j),

$$\sup_{y,z\in A'} |f(y;m) - f(z;m)| \leq C c_{i(j)} 2^{-\beta(q+S-j)}.$$

The known scales satisfy

$$\frac{c_{i(j)}}{a_q} \leq C(1+q-j)^\gamma, \qquad a_q \leq a_j \leq C_a c_{i(j)}, \qquad \sup_{u\geq 0}(1+u)^\gamma 2^{-\beta u} < \infty,$$

so the oscillation of f on A' is at most $C_{\mathrm{cal}} a_q 2^{-\beta S}$, which is at most $\eta a_q/2$ by the choice of R in (48) ($2^{-\beta S} \leq 2^{-\beta R} \leq \eta/(2C_{\mathrm{cal}})$). Combining with (60) and $\frac{\eta}{4} a_{q+S} \leq \frac{\eta}{4} a_q$ gives (62). The accuracy of the held score is the one of its own birth, at the scale of its query depth; no common final accuracy of all steps is used. This is a guarantee for the held score itself; no re-selection of the critic of step j is needed. □

Deep core mass and exponential moments. The calibration scale a_q varies over the response space, and the next lemma controls its distribution under P_{start}; the exponential moment is what bounds the density ratio in the regret transfer.

Lemma 26 (Normalization, deep-core mass and moments of the scale)**.** *Let $K = \rho^{-1/\gamma}$, $k_0 = \lceil K \rceil$, and let C_H be large enough that $H \geq \bar{\ell} k_0 + 1$. On $\mathcal{E}$, there are constants $U, C_{\exp}$, independent of $n, M, \rho, r_{\mathrm{mom}}$ and S, such that*

$$Z_H := \int e^{W_H/\rho}\, \mathrm{d}\mu \geq \exp\left(\frac{G_m}{\rho} - UK - C\eta\right), \tag{63}$$

$$P_{\mathrm{start}}(C_{\mathrm{core}} \mid m) + P_{f,\rho}(C_{\mathrm{core}} \mid m) \leq C \exp\left(-dH\log 2 + UK + C\eta\right) \leq \delta_C := \frac{\rho^2}{(n+1)^2}, \tag{64}$$

and, with $q = q(m,y)$ and for every fixed $\varpi \geq 0$,

$$\int \exp\left(\frac{r_{\mathrm{mom}}\eta a_q}{\rho}\right) \mathrm{d}P_{\mathrm{start}} \leq C_{\exp}, \qquad \int \left(\frac{a_q}{\rho}\right)^{\varpi} \mathrm{d}P_{\mathrm{start}} \leq C_\varpi, \tag{65}$$
$$\sum_{j<H} P_{\mathrm{start}}(A_j \mid m) a_j^2 \leq C\rho^2.$$

Proof. The three statements are proved in the order in which they depend on each other. We first lower-bound the partition function Z_H by restricting the integral to a near-optimal region on which the held score is accurate; this denominator bound, together with the tiny volume of the deepest region, shows that the deep core has negligible mass under both P_{start} and the true Gibbs policy. For the moments we split the remaining shells into two regimes: shallow shells have a large reward gap and hence exponentially small Gibbs weight, while deep shells may carry mass but there a_q/ρ is uniformly bounded. Combining the two regimes gives the exponential moment, the polynomial moments and the weighted quadratic bound.

Normalization. Let $\ell = L_{k_0} \leq \bar{\ell} k_0 \leq H-1$. Step ℓ is executed, and at this step the correct parent $P_\ell^*(m) = C_{k_0}(m)$ is selected (Lemma 20) and updated as a whole at depth $\ell+S$; since later updates only refine, $C_{k_0}(m)$ is a union of final cubes with last update index $q \geq \ell$, whose birth errors are at most $\frac{\eta}{4} a_{q+S} \leq \frac{\eta}{4} a_\ell \leq C\eta c_{k_0} \leq C\eta\rho$ by (60). By (18), $f \geq G_m - C_{\mathrm{near}} T_{k_0}$ on $C_{k_0}(m)$, and so are the cube averages of the final cubes inside; hence

$$W_H \geq G_m - C_{\mathrm{near}} T_{k_0} - C\eta\rho \ \text{ on } C_{k_0}(m),$$

and (63) follows from $\mu(C_{k_0}) \geq 2^{-d\bar{\ell}k_0}$, $k_0 \leq 2K$ and $T_{k_0}/\rho \leq C_T k_0^{-(\gamma-1)} K^\gamma \leq 2C_T K$, with $U = 4d\bar{\ell}\log 2 + 2C_{\mathrm{near}} C_T$. No constancy of the score on the core is assumed, and the core $C_{\mathrm{core}} = P_H^*(m)$ is also a union of final cubes, all created at step $H-1$.

Deep-core mass. The core has volume 2^{-dH}, and on it $q = H - 1$ and $W_H \le Q + \frac{\eta}{4} a_{H-1+S} \le G_m + C\eta\rho$ (as $a_{H-1+S} \le a_H \le C\rho$); together with (63) this gives the first inequality of (64) for P_{start}, and the same bound holds for $P_{f,\rho}$ by the lower bound on its normalization from (18). The second inequality holds once the constants C_H, C'_H in (48) are large.

Moments. Choose c_η with $2c_\eta C_a C_{cT} \le \kappa/2$, where $c_i \le C_{cT} T_i$. For a final cube $A' \subset A_j \subset E_i$ (exit level i), by (17), (62) and $a_q \le a_j \le C_a c_i$,

$$W_H + r_{\mathrm{mom}} \eta a_q \le G_m - \kappa T_i + (r_{\mathrm{mom}} + 1)\eta C_a c_i \le G_m - \frac{\kappa}{2} T_i,$$

and the same holds with a_q replaced by a_j. Fix $\theta_0 > 0$ with $\kappa c_T \theta_0^{1-\gamma}/2 \ge U + 2$. For $i \le \theta_0 K$, (63) and (16) give a contribution at most $C\mu(A')e^{-2K}$ (the factor $e^{C\eta} \le C$ is absorbed); for $i > \theta_0 K$, $a_q/\rho \le a_j/\rho \le C\theta_0^{-\gamma}$ and $r_{\mathrm{mom}}\eta = c_\eta$, so the exponential factor is bounded by a constant. Inside the core, $q = H - 1$ and $a_q \le CH^{-\gamma} \le C\rho$, so the core can be included in the first two bounds. Summing gives (65); the decomposition into shells is a device of the proof and does not correspond to any special treatment of the core by the algorithm. □

The final critic fit. Let Π be the joint law of $(X, D_{\mathrm{ref}}(X, Y))$ under P_{start} and $T_{\mathrm{ref}} = d(H + S + r) + 3$; in the rest of this subsection, N denotes the approximation size N_{ref} of (50), with $r = \lceil \alpha \log_2 N/(\beta d) \rceil$.

The final critic approximates a different target, the reward on each shell in the local coordinates of that shell, by the same pattern as in Appendix C.4.

Lemma 27 (Approximation and capacity for the final refinement)**.** *There is a class* $\mathcal{T}_{\mathrm{ref}} = \mathcal{T}(N_{\mathrm{blk}}, D_{\mathrm{emb}}, L_{\mathrm{FFN}}, W_{\mathrm{FFN}}, S_{\mathrm{FFN}}, B_{\mathrm{par}})$ *with* $T_{\mathrm{in}} = T_{\mathrm{ref}}$ *and*

$$\begin{aligned} D_{\mathrm{emb}} &= C_{\alpha,d}, \qquad N_{\mathrm{blk}} \le C_{\alpha,d}(R + 2 + \log(N+1)), \\ L_{\mathrm{FFN}} &\le C_{\alpha,d}(1 + \log(N+1)), \\ W_{\mathrm{FFN}}, S_{\mathrm{FFN}} &\le C_{\alpha,d} M(H+1)(N+1)(1+\log(N+1)), \\ B_{\mathrm{par}} &\le C_{\alpha,d}\big[M + T_{\mathrm{in}}^2 \log(8MT_{\mathrm{in}}) + 1\big], \end{aligned} \tag{66}$$

whose covering entropy satisfies

$$\log \mathcal{N}_\infty(\varepsilon, \mathcal{T}_{\mathrm{ref}}) \le CM(H+1)(N+1)(R+2+\log(N+1))^5 \log \frac{CMT_{\mathrm{in}}(N+1)}{\varepsilon}, \tag{67}$$

and which, on $\mathcal{E}$ *(with the front-phase history fixed as above), contains a comparator* F_0 *(depending on the history) with*

$$\|F_0 - Q\|^2_{L^2(\Pi)} \le C\rho^2 N^{-2\alpha/d} + 4\delta_C. \tag{68}$$

Consequently, the ERM $\widehat{T}_{\mathrm{ref}}$ *over* $\mathcal{T}_{\mathrm{ref}}$ *on* b *fresh records satisfies, for a constant* D_{E}*,*

$$\mathbb{E}\left[\left\| \widehat{T}_{\mathrm{ref}} - Q \right\|^2_{L^2(\Pi)} \,\middle|\, \mathcal{E} \right] \le C\rho^2 N^{-2\alpha/d} + 4\delta_C + \frac{C(1+\sigma^2)M(H+1)(N+1)}{b} \mathcal{L}^{D_{\mathrm{E}}}. \tag{69}$$

Proof. The proof repeats the approximation–realization–covering–ERM pattern of Appendix C.4 for the shell-wise target, in four steps. We first approximate the local Hölder function of each shell at resolution N by a ReLU network, keeping the side length of the shell out of the approximation problem; then we realize these networks inside the Transformer class with the fixed positional encoding; then we read off the capacity and the covering entropy; and finally we collect the approximation error, the contribution of the deep core and the estimation error of Lemma 15.

Local approximation without treating the side length as a variable. By Lemma 10, $h_{m,j}(z) = (g_{m,j}(z) - g_{m,j}(0))/a_j$ belongs to a fixed α-Hölder ball on $[0,1]^d$. By Schmidt-Hieber (2020, Theorem 5) (see also Yarotsky (2017)), there is a ReLU network $\tilde{h}_{m,j}$ of depth $C\log(N+1)$, width CN and $CN\log(N+1)$ nonzero weights of size $O(1)$ with

$$\left\| \tilde{h}_{m,j} - h_{m,j} \right\|_\infty \le CN^{-\alpha/d}$$

(for all $\alpha > 0$, including $C^{k,1}$ at integer α); the comparator uses $\tilde{g}_{m,j} := g_{m,j}(0) + a_j \tilde{h}_{m,j}$, so that $\|\tilde{g}_{m,j} - g_{m,j}\|_\infty \le C a_j N^{-\alpha/d}$, with MH local networks in parallel and $CMHN\log(N+1)$ nonzero weights in total. A query cube D of a response leaving the path at depth j has depth $q + S + r = q + R + 1 + r$ with $q \ge j$ by (61). Let $F = D_{j+R+1+r}(y)$ be its ancestor of depth $j + R + 1 + r$, and let z_F be the local coordinates of the center of F in P_j. Since $D \subset F \subset A_j$, Lemma 10 gives

$$\max\left\{|Q(m,D) - g_{m,j}(z_F)|,\ \sup_{y\in F}|f(y;m) - g_{m,j}(z_F)|\right\} \le C a_j 2^{-\beta(R+1+r)} \le C a_j N^{-\alpha/d}. \tag{70}$$

Here $2^{-\beta r} \le N^{-\alpha/d}$ by the choice of r. Thus only the first $d(R+1+r)$ bits after the exit from the path have to be read, whatever the query depth; no approximation in $d+1$ variables and no separate approximator per (m,j,q) is needed. Reading these $O_{\alpha,d}(R+1+\log(N+1))$ bits uses that many blocks, and does not require a table indexed by the 2^{dr} fine cubes; each cube query still costs one reward observation.

Realization with the fixed positional encoding. The comparator is built from the following blocks; D_{emb} is fixed throughout.

(1) *Reading the input.* The first block recovers from $\varphi(t)$, by Lemma 13 and plateaus, the token category and the integers t, t^2 at every position, with sparsity $O(T_{\mathrm{in}} + M)$ and weights $O(T_{\mathrm{in}}^2 + M)$; a second block delivers the prompt index m to all positions by an attention whose logit is $\log(16MT_{\mathrm{in}})$ on the prompt token, followed by a plateau on $[1, M]$.

(2) *Locating the exit from the correct path.* The comparator stores the dH bits of the correct path of each prompt in an FFN table of size $O(MH)$ — a property of the comparator, not information given to the learner — and flags as an *event* every one of the first dH response positions at which the input bit differs from the path bit, as well as the `STOP` position; later bits raise no event. With the event flag $e_t \in \{0,1\}$, the readout attends with the logit

$$b_0 e_t - a_0 t, \qquad a_0 = \log\bigl(16T_{\mathrm{in}}^2\bigr), \quad b_0 = a_0(T_{\mathrm{in}} + 1),$$

so that the first event position s_{first} receives weight at least $7/8$; its position and `STOP` flag are recovered exactly, and the exit depth

$$j = \left\lfloor \frac{s_{\mathrm{first}} - 2}{d} \right\rfloor$$

is computed by $O(T_{\mathrm{in}})$ ReLUs.

(3) *Reading the local coordinates.* If the first event is a mismatch, $d(R+1+r)$ further blocks read the bits at the positions

$$dj + 2,\ \ldots,\ dj + d(R+1+r) + 1$$

by an attention with logit $2a'st - a't^2$, $a' = \log(8T_{\mathrm{in}})$, and keys (t, t^2); accumulating them into d coordinates with weights $2^{-k'}$ and adding $2^{-(R+r+2)}$ yields $z_F \in [0,1]^d$. All these bits are present in the input, since $q \ge j$.

(4) *Selecting the local network.* The ReLU gate

$$(1 - |m - m_0| - |j - j_0|)_+$$

selects the local network $\tilde{g}_{m_0,j_0}(z_F)$, so that the approximation error is not multiplied by MH.

(5) *The core.* If the first event is `STOP`, i.e. there is no mismatch within the first dH bits, the comparator returns 0; on the actual query support of a good history these inputs are exactly the queries inside C_{core}.

Capacity and covering. The construction fits (66): the sparsity is dominated by $O(MT_{\mathrm{in}})$ for the position/path tables and $CMHN\log(N+1)$ for the local networks, the number of blocks is $O(R+1+r) = O(R+2+\log(N+1))$, and the weights are polynomial in M, T_{in}; no address of the final

holding cubes has to be memorized, so no extra factor H enters the capacity. Lemma 14 gives (67) (the exponent 5 is not optimized).

Errors. On the queries outside the core, by (70) and the third bound of (65) applied to every prompt (recall that $\Pi(m,\cdot)$ carries the factor $1/M$),

$$\sum_{m=1}^{M}\sum_{j<H}\Pi(m,A_j)\,Ca_j^2N^{-2\alpha/d}\le C\rho^2N^{-2\alpha/d};$$

on the core, $Q\in[-1,1]$ and the comparator returns 0, so the contribution is at most $4\Pi(\text{core})\le 4\delta_C$ by (64). This gives (68); the value 0 on the core is a property of the comparator, not a removal of these records from the ERM. For the resolution (50), $\log(N+1)+\log T_{\text{ref}}\le C\mathcal{L}$. Lemma 15 with $n_b=b$ and $V(b^{-2})$ bounded by (67) gives (69). □

The previous lemma bounds the error at the queried cubes, while the output policy is built from the clipped score on the whole space; the transfer uses Lemma 9 and the clipping widths of (6).

Lemma 28 (Projection error and the clipped score)**.** *On $\mathcal{E}$, the clipped score $v=W_{H+1}$ of line 6 of the final refinement step satisfies $v\in[-1,1]$, and, with $f^\dagger=f$ off the core and $f^\dagger=W_H$ on the core,*

$$\mathbb{E}\Big[\int\big|v-f^\dagger\big|^2\,\mathrm{d}P_{\text{start}}\;\Big|\;\mathcal{E}\Big]\le C\rho^2N^{-\frac{2\alpha}{d}}+\frac{C(1+\sigma^2)M(H+1)(N+1)}{b}\mathcal{L}^{D_{\mathrm{E}}}+C\delta_C.\quad(71)$$

Proof. The index q is constant on each final cube, and the query cubes are the r-fold subdivisions of the final cubes. Since P_{start} has constant density on each final cube, it is uniform on each query cube D, so that

$$\int_D\Big(\widehat{T}_{\text{ref}}(m,D)-f\Big)^2\,\mathrm{d}P_{\text{start}}=P_{\text{start}}(D)\Big(\widehat{T}_{\text{ref}}(m,D)-Q(m,D)\Big)^2+\int_D\big(Q(m,D)-f\big)^2\,\mathrm{d}P_{\text{start}}\,.$$

Off the core, the second term summed over the query cubes and the prompts is at most $C\rho^2N^{-2\alpha/d}$ by (70) and (65). Off the core, (62) shows that f and $Q(m,D)$ lie in the clipping interval $[W_H-\Delta,W_H+\Delta]$, so the clipping is a projection and does not increase the distance to $f^\dagger=f$. On the core, the same update is applied, $q=H-1$, and $|v-W_H|\le\rho+\eta a_{H-1}\le C\rho$, so the squared error is at most $C\rho^2P_{\text{start}}(\text{core})\le C\rho^2\delta_C$. Combining with (69) gives (71). The clipping interval contains W_H and both W_H and $\widehat{T}_{\text{ref}}$ lie in $[-1,1]$, so $v\in[-1,1]$. □

From squared error to KL-regularized regret. A small squared error does not by itself give a small regret, and it is worth seeing why before the transfer is carried out. Let $\Phi_\rho(w)=\rho\log\int e^{w/\rho}\,\mathrm{d}\mu$. For a Gibbs policy $P_v=P_{v,\rho}$, the regret against a reward f is the Bregman divergence of Φ_ρ,

$$\mathcal{R}_{\rho,f}(P_v)=\Phi_\rho(f)-\Phi_\rho(v)-D\Phi_\rho(v)[f-v],\qquad D^2\Phi_\rho(w)[e,e]=\frac{1}{\rho}\operatorname{Var}_{P_w}(e),$$

so that, by Taylor's formula with integral remainder,

$$\mathcal{R}_{\rho,f}(P_v)=\frac{1}{\rho}\int_0^1(1-s)\operatorname{Var}_{P_{w_s}}(f-v)\,\mathrm{d}s\,,\qquad w_s=(1-s)v+sf.\quad(72)$$

The variance is taken under the interpolated Gibbs policies, not under the collecting policy, and in general a small squared error under the collecting policy does not control (72). Indeed, take $f=0$, $P_{\text{start}}=\mu$, and $v_\rho=\mathbb{1}_{A_\rho}$ for a set A_ρ of measure $e^{-1/(2\rho)}$; then

$$\|v_\rho-f\|^2_{L^2(P_{\text{start}})}=e^{-1/(2\rho)}\ll\rho,\qquad P_{v_\rho}(A_\rho)\xrightarrow[\rho\to0]{}1,\qquad\mathcal{R}_{\rho,0}(P_{v_\rho})\xrightarrow[\rho\to0]{}\frac{1}{2}.$$

A rare overestimate on a region of tiny mass concentrates the policy there, and bounding the density ratio by the range of the scores alone would cost a factor $e^{2/\rho}$. The calibration of Lemma 25 and the clipping (6) avoid both.

The conversion uses the calibration and the exponential moment to bound the density ratio along the interpolation.

Lemma 29 (Regret of the clipped update)**.** *Let $\bar{u}$ be a known upper bound on the right-hand side of* (71) *with $\bar{u} \le 1$, and suppose $r_{\mathrm{mom}} \ge \max\{2, \log(4/\bar{u})\}$. Then, on $\mathcal{E}$,*

$$\mathbb{E}[\mathcal{R}_{\rho,f}(P_v) \mid \mathcal{E}] \le C\frac{\bar{u}}{\rho} + C\delta_C. \tag{73}$$

Proof. The proof proceeds in three steps. By (72) the regret is an integral of variances under the Gibbs policies interpolating between the clipped score and the reward, so the argument has to compare those policies with the collecting policy P_{start}. We first bound the density ratio along the interpolation by $Ce^{\eta a_q/\rho}$, using the calibration of the held score; then Hölder's inequality with the exponential moment of Lemma 26 converts the squared error under P_{start} into a regret bound against the modified reward $f^\dagger$; and finally the core is restored, at the price of its mass δ_C.

Density ratios along the interpolation. For any interpolation w between $f^\dagger$ and v, we have $|w - W_H| \le \rho + \eta a_q$ everywhere (on the core because $q = H-1$). By (65) and Jensen's inequality,

$$\frac{Z_w}{Z_H} \ge \exp\left(-1 - \eta\int \frac{a_q}{\rho}\,\mathrm{d}P_{\mathrm{start}}\right) \ge c, \qquad \frac{\mathrm{d}P_w}{\mathrm{d}P_{\mathrm{start}}} \le C\exp\left(\frac{\eta a_q}{\rho}\right),$$

and the same bound holds on the core since $a_{H-1} \le C\rho$; no constancy of the score on the core is used.

Regret against $f^\dagger$. By (72) and the density bound, with $E := \left|v - f^\dagger\right| \le 2$,

$$\mathcal{R}_{\rho,f^\dagger}(P_v) \le \frac{C}{\rho}\int e^{\eta a_q/\rho}E^2\,\mathrm{d}P_{\mathrm{start}}\,.$$

Hölder's inequality with exponents r_{mom} and $r_{\mathrm{mom}}/(r_{\mathrm{mom}}-1)$ and (65) give

$$\begin{aligned}\int e^{\eta a_q/\rho}E^2\,\mathrm{d}P_{\mathrm{start}} &\le \left(\int e^{r_{\mathrm{mom}}\eta a_q/\rho}\,\mathrm{d}P_{\mathrm{start}}\right)^{1/r_{\mathrm{mom}}}\left(\int E^{2r_{\mathrm{mom}}/(r_{\mathrm{mom}}-1)}\,\mathrm{d}P_{\mathrm{start}}\right)^{1-1/r_{\mathrm{mom}}}\\ &\le 4\,C_{\mathrm{exp}}^{1/r_{\mathrm{mom}}}\left(\int E^2\,\mathrm{d}P_{\mathrm{start}}\right)^{1-1/r_{\mathrm{mom}}},\end{aligned}$$

using $E^{2r_{\mathrm{mom}}/(r_{\mathrm{mom}}-1)} \le 4^{1/(r_{\mathrm{mom}}-1)}E^2$. Taking expectations, using concavity and (71),

$$\mathbb{E}\,\mathcal{R}_{\rho,f^\dagger}(P_v) \le \frac{C}{\rho}\bar{u}^{1-1/r_{\mathrm{mom}}} \le \frac{Ce}{\rho}\bar{u},$$

where the last step uses $\bar{u}^{-1/r_{\mathrm{mom}}} \le e$ by the choice of r_{mom}; the constant of the exponential moment is uniform in r_{mom}, so no polynomial factor is lost.

Restoring the true reward on the core. By the density bound and (64), $P_v(C_{\mathrm{core}}) + P_{f^\dagger,\rho}(C_{\mathrm{core}}) \le C\delta_C$. Since f and $f^\dagger$ agree off the core,

$$\Phi_\rho(f) - \Phi_\rho(f^\dagger) = \rho\log\frac{1 - P_{f^\dagger,\rho}(C_{\mathrm{core}})}{1 - P_{f,\rho}(C_{\mathrm{core}})},$$

whose absolute value is at most $C\rho\delta_C$, while the expected rewards of P_v under f and $f^\dagger$ differ by at most $2P_v(C_{\mathrm{core}}) \le C\delta_C$. This gives (73). □

Proof of Proposition 24. On $\mathcal{E}$, Lemmas 25 to 28 hold for every prompt, and (71) bounds the squared error of the clipped score v under P_{start} by

$$\bar{u}_{N_{\mathrm{ref}}} := C\rho^2 N_{\mathrm{ref}}^{-\frac{2\alpha}{d}} + \frac{C(1+\sigma^2)M(H+1)(N_{\mathrm{ref}}+1)}{b}\mathcal{L}^{D_{\mathrm{E}}} + C\delta_C.$$

For the choice (50), put $\mathsf{A} = b\rho^2/[(1+\sigma^2)M(H+1)\mathcal{L}^{D_{\mathrm{E}}}]$. The large-budget condition, $b \ge n/2$ and $H+1 \le C\rho^{-1/\gamma}\mathcal{L}$ give $\mathsf{A} \ge cC_{\mathrm{F}}M^{\zeta_M-1}\mathcal{L}^{D_{\mathrm{F}}-D_{\mathrm{E}}-1}$. Taking $C_{\mathrm{F}}, D_{\mathrm{F}}$ sufficiently large makes $\mathsf{A} \ge 1$, and $N_{\mathrm{ref}} \asymp \mathsf{A}^{d/(2\alpha+d)}$ yields $\bar{u}_{N_{\mathrm{ref}}} \le C\rho^2[\mathsf{A}^{-p} + (n+1)^{-2}] \le \rho^2 \le 1$. The known bound may be enlarged to at least $c\rho^2/(n+1)^2$, so that $\log(4/\bar{u}_{N_{\mathrm{ref}}}) \le C\mathcal{L} \le r_{\mathrm{mom}}$ for C_{mom} large and the clipping widths use only ρ, η and a_q. Lemma 29 then gives $\mathbb{E}[\mathcal{R}_\rho(\widehat{P}) \mid \mathcal{E}] \le C\bar{u}_{N_{\mathrm{ref}}}/\rho$, which is (59). □

D.4 Proof of the Large-Budget Bound

By Proposition 18, the front phase uses $\sum_{h<H} b_h \le n/2$ queries on the range $n \ge C_{\mathrm{F}}(1+\sigma^2)M^{\zeta_M}\rho^{-(2+1/\gamma)}\mathcal{L}^{D_{\mathrm{F}}}$ and the event $\mathcal{E}$ has probability at least $1-\delta_F$; the final iteration uses the remaining $b \ge n/2$ queries, so the total is at most n, and $H+1 \le C\rho^{-1/\gamma}\mathcal{L}$. The choice (50) of N_{ref} balances the first two terms of (59), so that Proposition 24 gives, on $\mathcal{E}$,

$$\mathbb{E}\Big[\mathcal{R}_\rho(\widehat{P}) \;\Big|\; \mathcal{E}\Big] \le C\rho\left[\frac{(1+\sigma^2)M(H+1)}{b\rho^2}\right]^p \mathcal{L}^{D_{\mathrm{E}}} + C\frac{\delta_C}{\rho},$$

while on $\mathcal{E}^c$ the regret is at most 4, since all scores lie in $[-1,1]$. Since $\Pr(\mathcal{E}^c) \le \delta_F = \rho^2/(n+1)^2$, $\delta_C/\rho = \rho/(n+1)^2$, $H+1 \le C\rho^{-1/\gamma}\mathcal{L}$ and $b \ge n/2$, and since $p<1$ and $n\rho^{2+1/\gamma}/M \le n$, the failure terms are absorbed into the main term, and

$$\mathbb{E}\Big[\mathcal{R}_\rho(\widehat{P})\Big] \le C\rho\left(\frac{M\rho^{-1/\gamma}}{n\rho^2}\right)^p \mathcal{L}^D = C\rho\left(\frac{M}{n\rho^{2+1/\gamma}}\right)^p \mathcal{L}^D,$$

which is (7); the factor $(1+\sigma^2)^p$ is absorbed into C. The dependence on M enters the starting condition through M^{ζ_M}, i.e., through the front-phase cost of Proposition 18, and the rate itself through the joint estimation error of the final fit, which carries the factor M of (69) only. □

E Proof of the Small-Regularization Bound of Theorem 3

In this section, Algorithm 1 runs with the small-regularization hyperparameters of Section 3.2: $\ell_{\mathrm{look}} = S = 1$, $H_{\mathrm{tot}} = H$ exploration iterations with the query rule (5) and $\Delta_h = 2$, no refinement step, and the output is the last actor $\widehat{P} = P_H = P_{W_H,\vartheta_H}$. The conventions of Appendix D are in force, with

$$t = \frac{n}{M}, \qquad L_n = \log(e+n+M), \qquad p_0 = \frac{\gamma-1}{2\gamma+1}, \qquad \mathsf{T}_h = \sum_{j\ge h} a_j \asymp (h+1)^{-(\gamma-1)},$$

and $\mathcal{L}$ replaced by L_n in the lemmas of Appendix D.2. The regularization ρ enters neither the algorithm nor the analysis until the last step of the proof; in particular $\rho = 0$ is allowed throughout. The lemmas of Appendices C.2 to C.4 are used as they stand, and the lemmas of Appendix D.2 are used with $S = 1$, so that a cube created at step r has depth $r+1$, the accuracy of step h is $\varepsilon_h = \frac{\eta}{4}a_{h+1}$, and $p_S = \kappa_0 2^{-d} =: p_{\mathrm{fine}}$.

Proof strategy. The proof has the same two stages as that of the large-budget bound, and differs only in the second. The front phase is the same loop, run with lookahead $S = 1$, and Appendix E.2 shows that its updates stay birth-accurate and that the budget fixes the depth H it reaches. There is no final refinement step: the output is the last actor, and Appendix E.3 converts the reached depth directly into a reward gap through the Gibbs variational principle. Appendix E.4 combines the two and absorbs the regularization penalty.

E.1 Hyperparameters of the Small-Regularization Setting

Let $\eta \le \eta_0$ be a fixed structural constant (small enough for Lemma 12 and Lemma 32), $\kappa_0 = p_*/2$ with p_* from Lemma 12, $\vartheta_h = t_0 c_{1+\lfloor h/\ell_\star \rfloor}$ for all $h \ge 0$ with t_0 as fixed after Lemma 12, and

$$J = \lceil C_J L_n \rceil, \qquad K_w = C_d(J+3), \qquad N = \left\lceil C_N M^{\frac{d}{2\alpha}} L_n^{D_N} \right\rceil, \tag{74}$$

$$b_h = \left\lceil C_b(1+\sigma^2)M^{\zeta_M} a_{h+1}^{-2} L_n^{D_b} \right\rceil, \qquad \bar{\delta}_{H'} = (H'+1)^{-\gamma}, \qquad \delta_h = \frac{\bar{\delta}_{H'}}{H'+1} \quad (h < H'),$$

for a candidate number of iterations $H' \ge 1$; the structural constants C_J, C_N, C_b, D_N, D_b are chosen as in Proposition 18. The number of iterations is the largest one that the budget allows,

$$H = \max\left\{ H' \in \{1,\dots,n\} : F_n(H') := \sum_{h<H'} b_h \le n \right\}, \tag{75}$$

and $H = 0$ if the set is empty (in particular if $n = 0$), in which case the uniform policy is returned without any query. The input length is $T_{\text{in}} = dH + 3$, and $\mathcal{T}_h = \mathcal{T}_{\text{win}}(K_w, N, T_{\text{in}})$ of Lemma 17, the same for all $h < H$. The total number of queries is $F_n(H) \le n$ deterministically, and the choice uses neither the reward, nor the zoom sequence, nor any per-prompt quantity.

E.2 The Front Phase

The front phase is the loop of Appendix D.2 run with $S = 1$: Lemma 30 shows that its updates stay birth-accurate, and Lemma 31 converts the budget into the depth it reaches.

Lemma 30 (Progress)**.** *With the parameters of Appendix E.1,* $\Pr(\mathcal{E}) \ge 1 - \bar{\delta}_H$*. On* $\mathcal{E}$*, for* $H \ge 1$*: the correct cube* $P^*_H(m)$ *is a union of cubes of* $\mathcal{D}_H(m)$ *for every* m*, all cubes of* $\mathcal{D}_H(m)$ *have depth at most* H*, and* $W_H(m, \cdot)$ *is* η*-birth-accurate on* $\mathcal{D}_H(m)$*.*

Proof. With $S = 1$, Lemmas 20 to 23 apply at every step $h < H$ on $\mathcal{E}_h$, with $p_S = p_{\text{fine}}$ and $\Lambda_h \le Ce^{-cJ}$. The factors η^{-1}, p_{fine}^{-1}, K_w, ℓ_N and $\log(2b_h/\delta_h)$ are bounded by fixed powers of L_n, while the factor a_{h+1}^{-2} is paid explicitly by b_h in (74). For the leakage, $H \le n$ gives $a_{H+1}^{-2} \le C(n+1)^{2\gamma}$, so choosing C_J large in $J = \lceil C_J L_n \rceil$ yields

$$\Lambda_h \le Ce^{-cJ} \le c\,\frac{p_{\text{fine}}\eta^2 a_{H+1}^2}{M} \le c\,\frac{p_{\text{fine}}\eta^2 a_{h+1}^2}{M}.$$

The approximation and estimation terms are controlled as in Proposition 18, with L_n in place of $\mathcal{L}$. Hence, conditionally on $\mathcal{E}_h$ the step succeeds with probability at least $1 - \delta_h$, and $\Pr(\mathcal{E}^c) \le \sum_{h<H} \delta_h \le \bar{\delta}_H$. The remaining claims are Lemmas 19 and 20 at $h = H$. □

Lemma 31 (Cost and the chosen depth)**.** *For* $1 \le H' \le n$*,*

$$c(1+\sigma^2)M^{\zeta_M} H'^{2\gamma+1} \le F_n(H') \le C(1+\sigma^2)M^{\zeta_M} H'^{2\gamma+1} L_n^{D_b}, \tag{76}$$

and the depth (75) *satisfies*

$$c\Big(1 + \frac{n}{(1+\sigma^2)M^{\zeta_M} L_n^{D_b}}\Big)^{\frac{1}{2\gamma+1}} \le H + 1 \le C\Big(1 + \frac{n}{M}\Big)^{\frac{1}{2\gamma+1}}. \tag{77}$$

Proof. Both inequalities of (76) follow from $a_{h+1}^{-2} \asymp (h+2)^{2\gamma}$ and $\sum_{h<H'}(h+2)^{2\gamma} \asymp H'^{2\gamma+1}$, the rounding being absorbed by the constants. *Lower bound on* H. Let $A_n = (1+\sigma^2)M^{\zeta_M} L_n^{D_b}$. If $n \ge C_1 A_n$ with C_1 large, then $H' = \lfloor a_0 (n/A_n)^{1/(2\gamma+1)} \rfloor$ with a_0 small satisfies $1 \le H' \le n$ and $F_n(H') \le n$ by (76), so $H \ge H'$; if $n < C_1 A_n$, the left inequality of (77) holds trivially with $c \le (1 + C_1)^{-1/(2\gamma+1)}$, including $H = 0$. *Upper bound.* If $H \ge 1$, then $n \ge F_n(H) \ge cM^{\zeta_M} H^{2\gamma+1} \ge cMH^{2\gamma+1}$, so $H \le C(n/M)^{1/(2\gamma+1)}$; the case $H = 0$ is trivial. □

The cost counts every record: stopped responses, responses of prompts that were never selected, and the batches after a failure. The factor M^{ζ_M}, rather than M, is the price of converting the joint risk of the critic into the accuracy on the update inputs of every prompt (Lemma 23); the depth reached by the algorithm therefore grows like $(n/M^{\zeta_M})^{1/(2\gamma+1)}$ up to logarithms, and not like $(n/M)^{1/(2\gamma+1)}$.

E.3 The Terminal Bound

Unlike the large-budget setting, no further critic is fitted: the output is the last actor and its regret is read off the depth it reached.

Lemma 32 turns the heterogeneous birth errors into a one-sided envelope of the score, and Lemma 33 combines that envelope with the Gibbs variational principle and the deterministic KL bound.

Recall that the output is $\widehat{P} = P_{W_H, \vartheta_H}$ with $\vartheta_H = t_0 c_{1+\lfloor H/\ell_\star \rfloor} \asymp a_H \asymp (H+1)^{-\gamma}$.

Lemma 32 (Envelope of a birth-accurate score)**.** *Let $\eta \le \eta_0$, fix m and $H \ge 1$, and let W be η-birth-accurate on a partition $\mathcal{D}$ of Ω into dyadic cubes of depth at most H such that $P_H^*(m)$ is a union of cubes of $\mathcal{D}$. Then, for all $y \in \Omega$,*

$$G_m - f(y;m) \le C\big[(G_m - W(y))_+ + \mathsf{T}_H\big], \tag{78}$$

$$W(y) \le G_m + Ca_H, \tag{79}$$

$$W(y) \ge G_m - C\mathsf{T}_H \qquad (y \in P_H^*(m)). \tag{80}$$

Proof. Let $k = k(H)$ and $y \in D \in \mathcal{D}$. By Lemma 8 (applied as in Lemma 12), either $D \subset E_a(m)$ for some $a < k$, or $D \subset C_{k-1}(m)$. *Shells.* If $D \subset E_a(m)$, then by (17) and (18), $\kappa T_a \le G_m - f(y';m) \le C_{\mathrm{near}} T_{a-1} \le CT_a$ for all $y' \in E_a(m)$, so that $G_m - Q(m,D) \ge \kappa T_a$; since $\mathrm{depth}(D) \ge L_{a-1} + 1$, the birth error is at most $\frac{\eta}{4} a_{\mathrm{depth}(D)} \le C\eta c_a \le C\eta T_a$, and for η small

$$G_m - W(D) \ge \frac{\kappa}{2} T_a > 0, \qquad G_m - f(y;m) \le CT_a \le C(G_m - W(D)) = C(G_m - W(y))_+,$$

which gives (78) and (79) on the shells. *Inside.* If $D \subset C_{k-1}(m)$, then $\mathrm{depth}(D) \ge L_{k-1} + 1$ and the birth error is at most $C\eta c_k \le Ca_H$, while $0 \le G_m - f(y';m) \le C_{\mathrm{near}} T_{k-1} \le C\mathsf{T}_H$ for $y' \in C_{k-1}(m)$ by (18) and $T_{k-1} \asymp \mathsf{T}_H$; hence $G_m - f(y;m) \le C\mathsf{T}_H$, $W(D) \le Q(m,D) + Ca_H \le G_m + Ca_H$, and, if $D \subset P_H^*(m) \subset C_{k-1}(m)$, $W(D) \ge Q(m,D) - Ca_H \ge G_m - C\mathsf{T}_H - Ca_H \ge G_m - C\mathsf{T}_H$. □

Lemma 33 (Terminal reward and KL)**.** *For every history and every m,*

$$\mathrm{KL}\left(\widehat{P}(\cdot \mid m) \,\middle\|\, \mu\right) \le dH \log 2. \tag{81}$$

On $\mathcal{E}$, for every m,

$$G_m - \mathbb{E}_{\widehat{P}(\cdot|m)}[f(\cdot;m)] \le C\mathsf{T}_H \le C(H+1)^{-(\gamma-1)}. \tag{82}$$

Proof. KL. By Lemma 19, $\widehat{P}(\cdot \mid m)$ is uniform inside every cube of $\mathcal{D}_H(m)$, whose cubes have depth at most H (with $S = 1$); refining $\mathcal{D}_H(m)$ to the uniform grid of depth H keeps $\widehat{P}$ uniform on each cell, so with $p_D = \widehat{P}(D \mid m)$ for the 2^{dH} cells,

$$\mathrm{KL}\left(\widehat{P}(\cdot \mid m) \,\middle\|\, \mu\right) = \sum_D p_D \log \frac{p_D}{2^{-dH}} = dH \log 2 + \sum_D p_D \log p_D \le dH \log 2.$$

This holds on every history; for $H = 0$ the divergence is 0.

Reward. For $H = 0$ the claim is trivial since $G_m - \mathbb{E}_\mu f \le 2$; let $H \ge 1$ and write $W = W_H(m,\cdot)$, which is η-birth-accurate on $\mathcal{D}_H(m)$ on $\mathcal{E}$ by Lemma 30. Let $U_* = \mu(\cdot \mid P_H^*(m))$ be the uniform distribution on the correct cube, a union of cubes of $\mathcal{D}_H(m)$. The Gibbs variational principle for $\widehat{P} = P_{W,\vartheta_H}$, (80) and $\mathrm{KL}\,(U_* \,\|\, \mu) = dH \log 2$ give

$$\mathbb{E}_{\widehat{P}}[W] - \vartheta_H \mathrm{KL}\left(\widehat{P} \,\middle\|\, \mu\right) \ge \mathbb{E}_{U_*}[W] - \vartheta_H \mathrm{KL}\,(U_* \,\|\, \mu) \ge G_m - C\mathsf{T}_H - \vartheta_H dH \log 2,$$

hence, by $\mathrm{KL}\left(\widehat{P} \,\middle\|\, \mu\right) \ge 0$ and $\vartheta_H H \le Ca_H H \le C\mathsf{T}_H$,

$$\mathbb{E}_{\widehat{P}}[G_m - W] \le C\mathsf{T}_H.$$

By (79), $(G_m - W)_+ \le (G_m - W) + Ca_H$, so $\mathbb{E}_{\widehat{P}}[(G_m - W)_+] \le C\mathsf{T}_H$, and (78) yields (82). No sup-norm accuracy of the final score is used: the old cubes carry only the accuracy of their own birth, and the envelope of Lemma 32 is what the variational comparison needs. □

E.4 Proof of the Small-Regularization Bound

The number of queries is $F_n(H) \le n$ by (75). For every m, $J_\rho^*(m) = \rho \log \int e^{f(y;m)/\rho} \,\mathrm{d}\mu\,(y) \le G_m$ (and $J_0^*(m) = G_m$), so, by (81),

$$\begin{aligned}\mathcal{R}_\rho(\widehat{P}) &= \frac{1}{M} \sum_m \Big[J_\rho^*(m) - \mathbb{E}_{\widehat{P}(\cdot|m)}[f(\cdot;m)] + \rho \mathrm{KL}\left(\widehat{P}(\cdot \mid m) \,\middle\|\, \mu\right)\Big] \\ &\le \frac{1}{M} \sum_m \Big[G_m - \mathbb{E}_{\widehat{P}(\cdot|m)}[f(\cdot;m)]\Big] + \rho\, dH \log 2.\end{aligned}$$

On $\mathcal{E}$ the bracket is at most $C(H+1)^{-(\gamma-1)}$ by (82), and on $\mathcal{E}^c$ it is at most 2 since $f \in [-1,1]$; $\Pr(\mathcal{E}^c) \le \bar{\delta}_H = (H+1)^{-\gamma}$ by Lemma 30. Therefore

$$\mathbb{E}\Big[\mathcal{R}_\rho(\widehat{P})\Big] \le C(H+1)^{-(\gamma-1)} + \rho\, dH \log 2, \tag{83}$$

and for $\rho = 0$ the last term vanishes. For $0 < \rho \le c_{\mathrm{S}}(1+t)^{-\gamma/(2\gamma+1)}$, the upper bound in (77) gives

$$\begin{aligned}\rho\, dH \log 2 &\le C\rho(H+1)^{\gamma}(H+1)^{-(\gamma-1)} \\ &\le C c_{\mathrm{S}}(1+t)^{-\frac{\gamma}{2\gamma+1}}(1+t)^{\frac{\gamma}{2\gamma+1}}(H+1)^{-(\gamma-1)} \le C(H+1)^{-(\gamma-1)}.\end{aligned}$$

In both cases, the lower bound in (77) yields

$$\mathbb{E}\Big[\mathcal{R}_\rho(\widehat{P})\Big] \le C(H+1)^{-(\gamma-1)} \le C\left(1 + \frac{n}{M^{\zeta_M} L_n^{D_b}}\right)^{-p_0},$$

where the factor $(1+\sigma^2)$ is absorbed into C; this is the first case of (7), and the case $H = 0$ (uniform output, regret at most 2) is included by adjusting the constant. The constant c_{S} is a structural constant that may be taken as small as we like; on the range of the small-regularization bound we have $(1+t)\rho^{2+1/\gamma} \le c_{\mathrm{S}}^{2+1/\gamma}$, so the first lower bound of (11) applies on that whole range as soon as $c_{\mathrm{S}}^{2+1/\gamma} \le c_0$, where c_0 is the constant of the small-regularization bound of Theorem 4. □

F Proof of Theorem 4

Throughout this section, $\sigma > 0$ is fixed and $\lambda \ge \Lambda_{\alpha,d}$. In the lower bound we use independent $N(0,\sigma^2)$ noise, which belongs to $\mathcal{N}_\sigma$, and we fix the zoom depths $\ell_i \equiv \ell_\star$; we may even reveal them to the learner, which only strengthens the statement. We write

$$t = \frac{n}{M}, \qquad p = \frac{2\alpha}{2\alpha+d}, \qquad p_0 = \frac{\gamma-1}{2\gamma+1}, \qquad q = 2^{-\ell_\star}.$$

Both bounds are first proved for a single prompt and then transferred to M prompts; the transfer is Lemma 34 of Appendix F.1 and is common to the two. The rest of the section proves them one after the other, each in three steps: a hard family, the information bound for that family, and the assembly. The small-regularization bound comes first. Its family, built in Appendix F.2, places the bump of each level in one of two cubes; Appendix F.3 shows that a cube query at any depth carries little information about those placements, so that a constant fraction of them stays undecodable within the budget, and Appendix F.4 assembles the bound. The large-budget bound follows. Its family, built in Appendix F.5, has a Gibbs policy that spreads a comparable mass over a band of levels around $\rho^{-1/\gamma}$ and perturbs each level by bumps with random signs; Appendix F.6 bounds the information that the queries carry about those signs, and Appendix F.7 assembles the bound.

F.1 Reduction to a Single Prompt

For a fixed prompt, the problem is that of a single sequence $(g_i)_{i\ge1}$ with regret $J_\rho^* - J_\rho(P)$. For a prior π on single-prompt instances, let $B_\pi(k,\rho)$ be the Bayes risk of the best algorithm that uses k adaptive cube queries; it is nonincreasing in k.

Lemma 34 (Bayes transfer)**.** *Let $k = \max\{1, \lceil 2t \rceil\}$, let π be any prior on single-prompt instances that may depend on (n, M, ρ) but not on the data, and put the product prior $\pi^{\otimes M}$ on the M sequences. Then*

$$R^*(n, M, \rho) \ge \frac{1}{2} B_\pi(k, \rho). \tag{84}$$

Proof. Consider the strengthened experiment in which the learner is told the whole prompt sequence $X_1, \dots, X_n$ in advance and, when evaluated on prompt m, the true functions of all other prompts. Since the sequences are independent under the product prior and the noise is independent across rounds, the answers to queries on other prompts can be simulated from the known functions and independent noise; hence the only source of information about prompt m is its own

$$N_m = \sum_{t'=1}^{n} \mathbb{1}_{X_{t'}=m} \sim \mathrm{Bin}(n, 1/M)$$

queries, and adaptivity across prompts does not help. The Bayes regret on prompt m is therefore at least

$$\mathbb{E}_{N_m} B_\pi(N_m, \rho) \geq \Pr(N_m \leq k) B_\pi(k, \rho) \geq \frac{1}{2} B_\pi(k, \rho),$$

where the last step is Markov's inequality with $\mathbb{E}N_m = t \leq k/2$. Averaging over m and using that the minimax risk dominates the Bayes risk gives (84). The prior is fixed before the data and is not re-chosen according to the realized N_m. □

Lemma 35 (KL divergence of adaptive cube queries)**.** *Let f, f' be two instances and let $\bar{P}_f, \bar{P}_{f'}$ be the laws of the transcript (internal randomness, prompts, queried cubes, observations) of an algorithm with k queries under Gaussian noise. Then*

$$\mathrm{KL}\left(\bar{P}_f \,\middle\|\, \bar{P}_{f'}\right) = \frac{1}{2\sigma^2} \mathbb{E}_f \sum_{r'=1}^{k} \left[Q_f(X_{r'}, D_{r'}) - Q_{f'}(X_{r'}, D_{r'})\right]^2. \tag{85}$$

Proof. By the chain rule, the kernels of the internal randomness, of the prompt and of the choice of $D_{r'}$ given the past are identical under both laws and contribute zero; the observation kernel given $(X_{r'}, D_{r'})$ is Gaussian with the same variance, and the KL divergence of two such Gaussians is the squared mean difference divided by $2\sigma^2$. □

F.2 A Two-Choice Family

For $r_c = 1/8$ and $m_\alpha = \lceil \alpha \rceil + 1$, define on $\mathbb{R}^d$

$$\psi_0(u) = \begin{cases} [1 - (\|u\|_\infty / r_c)^\alpha]_+, & 0 < \alpha \leq 1, \\ [1 - \|u\|_2^2 / r_c^2]_+^{m_\alpha}, & \alpha > 1, \end{cases} \qquad \psi_z(u) = \psi_0(u - z),$$

$$B_{\alpha,d} = \max\{8^\beta, |\psi_0|_\alpha\}, \qquad \Lambda_{\alpha,d} = 2B_{\alpha,d}. \tag{86}$$

Here $|\cdot|_\alpha$ is the zero-extension seminorm in **(A1)**. For $\alpha \leq 1$, $B_{\alpha,d} = 8^\alpha$, so $\Lambda_{\alpha,d} = 2 \cdot 8^\alpha$ as before. For $\alpha > 1$, the truncated polynomial is $C^{m_\alpha - 1}$ with bounded derivatives, hence $B_{\alpha,d} < \infty$, including integer α. In both cases ψ_0 has height one, a unique maximum at zero, and support inside $[-r_c, r_c]^d$; its value is strictly decreasing with radius inside its support. We use the following elementary gluing bound: disjoint translates of ψ_0, with either sign, have seminorm at most $2B_{\alpha,d}$. Indeed, at each point only one component contributes to any derivative; for the highest derivative at points in different supports, insert a zero value at each support boundary and bound each distance by the distance between the two points. The same argument applies to disjoint copies $A\psi_0((u-v)/h)$ with $h \leq 1$ and $|A|h^{-\alpha} \leq 1$: every derivative of order $j \leq \lceil \alpha \rceil - 1$ is bounded by the corresponding derivative of ψ_0, since $|A|h^{-j} \leq 1$, and the highest derivative has the same Hölder bound before gluing. Let $r_c = 1/8$, $q = 2^{-\ell_\star}$, $\ell_i \equiv \ell_\star$, and

$$\begin{aligned} z_0 &= (1/4 + q/2,\ 1/2 + q/2, \ldots, 1/2 + q/2), \\ z_1 &= (3/4 + q/2,\ 1/2 + q/2, \ldots, 1/2 + q/2), \\ z_- &= (1/2 + q/2,\ 1/2 + q/2, \ldots, 1/2 + q/2). \end{aligned}$$

For bits $\omega_{i,m} \in \{0, 1\}$ let

$$g_{1,m} = \psi_{z_{\omega_{1,m}}}, \qquad g_{i,m} = \psi_{z_{\omega_{i,m}}} - \psi_{z_-} \quad (i \geq 2). \tag{87}$$

Let Q_b be the depth-$\ell_\star$ dyadic cube centered at z_b, for $b \in \{0, 1\}$. Explicitly,

$$\begin{aligned} Q_0 &= [1/4, 1/4 + q) \times [1/2, 1/2 + q)^{d-1}, \\ Q_1 &= [3/4, 3/4 + q) \times [1/2, 1/2 + q)^{d-1}. \end{aligned}$$

The bits ω index reward instances, not response bits: $\omega_{i,m} = b$ selects a rescaled copy of Q_b inside $C_{i-1}(m)$ as the next active cell. The same disjoint local pair is used at every level and does not cover Ω. By **(A3)**, $q < r_c$. The bumps have disjoint supports and equal integrals, so the deeper primitives have zero mean; their zero extensions have $|g_i^0|_\alpha \leq \Lambda_{\alpha,d} \leq \lambda$ by the gluing bound. Each primitive has maximum one at $z_{\omega_{i,m}}$, so the family lies in $\mathcal{F}_{M,\theta}$ whenever $\lambda \geq \Lambda_{\alpha,d}$. The adaptive lower bound of Appendix F.3 and the fixed-sampling bounds of Appendix G both use this family.

F.3 Small Regularization: The Adaptive Exploration Lower Bound

By Lemma 34 and $1 + k \le 3(1 + t)$, it suffices to exhibit, for a single prompt, a prior $\pi_{k,0}$ with

$$B_{\pi_{k,0}}(k, 0) \ge c(1 + k)^{-p_0}, \tag{88}$$

which Lemma 36 then transfers from $\rho = 0$ to $\rho > 0$.

Lemma 36 (The regularization changes the optimal value by $O(\rho^{1-1/\gamma})$)**.** *There is a constant U depending only on θ such that for all $f \in \mathcal{F}_{M,\theta}$, $m \in [M]$ and $\rho \in (0, 1]$,*

$$0 \le G_m - J^*_\rho(m) \le U\rho^{1-1/\gamma}. \tag{89}$$

Consequently, for every policy P, $\mathcal{R}_\rho(P) \ge \mathcal{R}_0(P) - U\rho^{1-1/\gamma}$, and

$$R^*(n, M, \rho) \ge \Big[R^*(n, M, 0) - U\rho^{1-1/\gamma}\Big]_+. \tag{90}$$

Proof. Since $f \le G_m$, $J^*_\rho(m) = \rho \log \int e^{f/\rho} \,\mathrm{d}\mu \le G_m$. Let $K' = \lceil \rho^{-1/\gamma} \rceil$ and let $U_{K'} = \mu(\cdot \mid C_{K'}(m))$; by the variational formula, (18), (16) and $L_{K'} \le K'\bar{\ell} \le 2\bar{\ell}\rho^{-1/\gamma}$,

$$\begin{aligned} J^*_\rho(m) &\ge \mathbb{E}_{U_{K'}} f - \rho \mathrm{KL}\left(U_{K'} \parallel \mu\right) \ge G_m - C_{\mathrm{near}} T_{K'} - \rho\, d L_{K'} \log 2 \\ &\ge G_m - \big[C_{\mathrm{near}} C_T + 2d\bar{\ell} \log 2\big]\rho^{1-1/\gamma}. \end{aligned}$$

For a policy P, $\mathcal{R}_\rho(P) = \frac{1}{M}\sum_m [J^*_\rho(m) - \mathbb{E}_P f + \rho \mathrm{KL}\left(P \parallel \mu\right)] \ge \frac{1}{M}\sum_m [G_m - \mathbb{E}_P f] - U\rho^{1-1/\gamma} = \mathcal{R}_0(P) - U\rho^{1-1/\gamma}$. Taking the supremum over instances and the infimum over algorithms (an algorithm for the problem with parameter ρ is an algorithm for $\rho = 0$ with the same queries) gives (90), together with $R^* \ge 0$. □

Lemma 37 (Bayes lower bound for adaptive exploration of one prompt)**.** *Let $M = 1$ and $\rho = 0$. For every integer $s \ge 2$ there is a prior π_s on the two-choice family* (87) *such that every algorithm with $k \ge 0$ adaptive cube queries has Bayes regret*

$$B_{\pi_s}(k, 0) \ge \frac{\kappa T_{2s-1}}{4}\left[1 - \frac{C_{\mathrm{adv}}\, k\, c_s^2}{8\sigma^2 s}\right]_+^2, \qquad C_{\mathrm{adv}} = 8\sum_{r \ge 1} r^2 2^{-r/2}. \tag{91}$$

Consequently, with $s_k = \max\{2, \lceil (C_{\mathrm{adv}} k/(4\sigma^2 \zeta(\gamma)^2))^{1/(2\gamma+1)} \rceil\}$,

$$B_{\pi_{s_k}}(k, 0) \ge c(1 + k)^{-p_0}. \tag{92}$$

Proof. We first prove the finite-s bound by placing a prior on the unknown placements and pairing instances whose words to be decoded differ. For each fixed response, the reward has a spine structure that controls the information in any cube query. A Hellinger-affinity potential carries this control through adaptive histories; decoding the output policy then turns the remaining uncertainty into regret. The stated choice of s yields the rate.

Prior and comparison kernel. Let $\pi = \pi_s$ be the law of the reward-instance bits $\omega = (\omega_i)_{i \ge 1}$, where

$$\omega_i = 0 \quad (i < s), \qquad (\omega_i)_{i \ge s} \text{ are independent fair bits.}$$

The corresponding reward is f_ω from (87); by Appendix F.2, every f_ω lies in the class. The word to decode and its index set are

$$\varpi(\omega) = (\omega_s, \dots, \omega_{2s-1}), \qquad \mathcal{I}_s = \{s, \dots, 2s - 1\}.$$

For $i \in \mathcal{I}_s$, define Γ_i on $L^2(\pi)$ by keeping the bits before i, flipping ω_i, and resampling the bits after i. This is an analytical comparison between reward instances, not an operation performed by the learner. Its kernel density with respect to π is

$$\mathcal{K}_i(\omega, \omega') = 2^{i-s+1} \mathbb{1}_{\omega_{s:i-1} = \omega'_{s:i-1},\, \omega_i \ne \omega'_i}.$$

It is symmetric, nonnegative, and has unit row integrals, so $\|\Gamma_i\|_{L^2(\pi) \to L^2(\pi)} \le 1$. Set

$$\Gamma = \sum_{i \in \mathcal{I}_s} \Gamma_i, \qquad \mathcal{K} = \sum_{i \in \mathcal{I}_s} \mathcal{K}_i.$$

Then $\Gamma\mathbf{1} = s\mathbf{1}$ and $\|\Gamma\| \le s$. Moreover, $\mathcal{K}(\omega,\omega') > 0$ implies $\varpi(\omega) \ne \varpi(\omega')$.

The reward at a fixed point. Fix $y \in \Omega$. We claim that there are bits $\mathsf{a} = (\mathsf{a}_s, \mathsf{a}_{s+1}, \dots)$ and reals $(F_j)_{j\ge s}$ such that, with

$$\mathcal{B}_j = \{\omega : \omega_{s:j-1} = \mathsf{a}_{s:j-1},\ \omega_j \ne \mathsf{a}_j\}, \qquad \pi(\mathcal{B}_j) = 2^{-(j-s+1)}, \qquad j \ge s, \tag{93}$$

which partition π-almost all of the space,

$$f_\omega(y) = F_j\ \ (\omega \in \mathcal{B}_j), \qquad |F_{j+1} - F_j| \le 2c_j, \qquad \text{so} \quad |F_j - F_i| \le 2c_s(j-i)\ \ (s \le i < j). \tag{94}$$

The levels before s and the cell C_{s-1} do not depend on ω. If $y \notin C_{s-1}$, all later components vanish at y. Thus $f_\omega(y) = \sum_{i<s} c_i h_i(y)$ for every ω; choose any a and set every F_j to this common value.

Suppose now that $y \in C_{s-1}$. Choose a recursively: given $\mathsf{a}_{s:i-1}$ with $y \in C_{i-1}(\mathsf{a}_{<i})$, let w be the local coordinate of y in that cell. There are two cases because Q_0 and Q_1 do not cover Ω.

If $w \in Q_b$ for $b \in \{0,1\}$, set $\mathsf{a}_i = b$. For $\omega \in \mathcal{B}_i$ we have $\omega_i = 1-b$, so $y \notin C_i(\omega)$ and all components after level i vanish at y. Also, w is at $\|\cdot\|_\infty$-distance greater than r_c from both z_{1-b} and z_-, so the level-i component vanishes in this two-choice family. Therefore

$$f_\omega(y) = \sum_{l<i} c_l h_l(y) =: F_i \qquad (\omega \in \mathcal{B}_i).$$

The earlier reward components in this sum need not vanish.

If $w \notin Q_0 \cup Q_1$, then $y \notin C_i(\omega)$ for either choice of ω_i. All components after level i vanish, while the level-i component may be nonzero:

$$f_\omega(y) = \sum_{l<i} c_l h_l(y) + c_i\big[\psi_{z_{\omega_i}}(w) - \psi_{z_-}(w)\big].$$

Set $\mathsf{a}_i = 0$ and let F_i be the value for $\omega_i = 1$. For $j > i$, set $\mathsf{a}_j = 0$ and let F_j be the value for $\omega_i = 0$; the recursion stops because later bits have no effect on the reward at y.

When the recursion continues through level i, moving the first mismatch from i to $i+1$ changes the level-i contribution by at most c_i, and may add a level-$(i+1)$ contribution of at most c_{i+1} if the recursion stops there. At the stopping level the difference is at most

$$c_i|\psi_{z_0}(w) - \psi_{z_1}(w)| \le c_i;$$

thereafter the values F_j are constant. The bumps have disjoint supports and height at most one, so in all cases $|F_{j+1} - F_j| \le c_j + c_{j+1} \le 2c_j$. Summing these differences and using $c_l \le c_s$ for $l \ge s$ proves the last bound in (94).

Information of one cube query. Let T_y be the operator on $L^2(\pi)$ with kernel

$$\mathcal{K}(\omega,\omega')\big(f_\omega(y) - f_{\omega'}(y)\big)^2,$$

and normalize the indicators of the sets in (93) by

$$\mathsf{b}_j = \frac{\mathbb{1}_{\mathcal{B}_j}}{\sqrt{\pi(\mathcal{B}_j)}}.$$

Take $\omega \in \mathcal{B}_i$ and $\omega' \in \mathcal{B}_j$ with $i < j$. Their bits agree before i and differ at i, so only $\mathcal{K}_i$ can be nonzero; this requires $i \in \mathcal{I}_s$. On $\mathcal{B}_j \times \mathcal{B}_j$ the reward difference is zero. Thus the kernel of T_y is constant on each block $\mathcal{B}_i \times \mathcal{B}_j$, and the operator vanishes outside the closed span of $(\mathsf{b}_j)_{j\ge s}$. For $i < j$ its matrix entries are

$$\langle \mathsf{b}_i, \mathsf{T}_y \mathsf{b}_j\rangle = \mathbb{1}_{i\in\mathcal{I}_s}\, 2^{i-s+1}(F_j - F_i)^2\sqrt{\pi(\mathcal{B}_i)\pi(\mathcal{B}_j)} = \mathbb{1}_{i\in\mathcal{I}_s}\, 2^{-(j-i)/2}(F_j - F_i)^2. \tag{95}$$

By (94),

$$(F_j - F_i)^2 \le 4c_s^2(j-i)^2.$$

Each row of this symmetric matrix therefore has absolute sum at most

$$8c_s^2 \sum_{r\ge 1} r^2 2^{-r/2} = C_{\mathrm{adv}} c_s^2.$$

The Schur test gives $\|\mathsf{T}_y\| \le C_{\mathrm{adv}} c_s^2$. For a dyadic cube D and a nonnegative $v \in L^2(\pi)$, Jensen's inequality applied to the cube average and $\mathcal{K} \ge 0$ give

$$\begin{aligned}\iint \mathcal{K}(\omega,\omega')v(\omega)v(\omega')(Q_\omega(D)-Q_{\omega'}(D))^2 \,\mathrm{d}\pi(\omega)\,\mathrm{d}\pi(\omega') &\le \frac{1}{\mu(D)}\int_D \langle v, \mathsf{T}_y v\rangle \,\mathrm{d}\mu(y)\\ &\le C_{\mathrm{adv}} c_s^2 \|v\|_{L^2(\pi)}^2, \end{aligned} \tag{96}$$

uniformly in the depth of D.

A potential along the adaptive history. Fix an algorithm and use $N(0,\sigma^2)$ observation noise. Let $P_{\omega,t}$ be the law under f_ω of the history h_t: the internal randomness, queried cubes, and observations through query t. Let $P_{\mathrm{ref},t}$ be the law of the same algorithm when every observation is $N(0,\sigma^2)$ regardless of the query. This is a comparison law and need not be a reward instance. The Gaussian observation densities are positive, so the likelihood ratio exists and satisfies

$$\mathsf{L}_t(\omega,\mathsf{h}) = \frac{\mathrm{d}P_{\omega,t}}{\mathrm{d}P_{\mathrm{ref},t}}(\mathsf{h}), \qquad \int \mathsf{L}_t(\omega,\mathsf{h})\,\mathrm{d}P_{\mathrm{ref},t}(\mathsf{h}) = 1.$$

For probability laws P, P' dominated by a common law R, write their Hellinger affinity as

$$\mathrm{Aff}(P,P') := \int \sqrt{\frac{\mathrm{d}P}{\mathrm{d}R}\frac{\mathrm{d}P'}{\mathrm{d}R}}\;\mathrm{d}R\,.$$

Put $v_{\mathsf{h}}(\omega) = \sqrt{\mathsf{L}_t(\omega,\mathsf{h})}$ and define

$$\begin{aligned}\Phi_t &= \int \langle v_{\mathsf{h}}, \Gamma v_{\mathsf{h}}\rangle \,\mathrm{d}P_{\mathrm{ref},t}(\mathsf{h}),\\ \Phi_0 &= \langle \mathbf{1}, \Gamma\mathbf{1}\rangle = s, \qquad \int \|v_{\mathsf{h}}\|_{L^2(\pi)}^2 \,\mathrm{d}P_{\mathrm{ref},t}(\mathsf{h}) = 1.\end{aligned}$$

Expanding Γ with $R = P_{\mathrm{ref},t}$ shows what this potential measures:

$$\begin{aligned}\Phi_t &= \iint \mathcal{K}(\omega,\omega')\left[\int \sqrt{\mathsf{L}_t(\omega,\mathsf{h})\mathsf{L}_t(\omega',\mathsf{h})}\,\mathrm{d}P_{\mathrm{ref},t}(\mathsf{h})\right]\mathrm{d}\pi(\omega)\,\mathrm{d}\pi(\omega')\\ &= \iint \mathcal{K}(\omega,\omega')\,\mathrm{Aff}(P_{\omega,t},P_{\omega',t})\,\mathrm{d}\pi(\omega)\,\mathrm{d}\pi(\omega').\end{aligned}$$

The affinity equals one for identical laws and becomes smaller as they separate; Φ_t sums these affinities over pairs selected by $\mathcal{K}$.

Condition on the current history and any fresh algorithmic randomness used to choose the next cube. The resulting cube D_{t+1} is then the same under every reward instance. For arbitrary $u, u' \in \mathbb{R}$, completing the square gives

$$\begin{aligned}\mathrm{Aff}\big(N(u,\sigma^2),N(u',\sigma^2)\big) &= \int_{\mathbb{R}} \frac{\exp\big(-[(z-u)^2+(z-u')^2]/(4\sigma^2)\big)}{\sqrt{2\pi\sigma^2}}\;\mathrm{d}z\\ &= \exp\left[-\frac{(u-u')^2}{8\sigma^2}\right]\int_{\mathbb{R}} \frac{\exp\big(-(z-(u+u')/2)^2/(2\sigma^2)\big)}{\sqrt{2\pi\sigma^2}}\;\mathrm{d}z\\ &= \exp\left[-\frac{(u-u')^2}{8\sigma^2}\right].\end{aligned}$$

Thus, for the next query,

$$\mathrm{Aff}\big(N(Q_\omega(D_{t+1}),\sigma^2),N(Q_{\omega'}(D_{t+1}),\sigma^2)\big) = \exp\left[-\frac{\big(Q_\omega(D_{t+1})-Q_{\omega'}(D_{t+1})\big)^2}{8\sigma^2}\right].$$

Therefore $1 - e^{-z} \le z$ and (96) yield

$$0 \le \Phi_t - \Phi_{t+1} = \int \iiint \mathcal{K}(\omega, \omega') v_{\mathsf{h}}(\omega) v_{\mathsf{h}}(\omega') \Big[1 - e^{-(Q_\omega(D_{t+1}) - Q_{\omega'}(D_{t+1}))^2/(8\sigma^2)}\Big] \, \mathrm{d}\pi \, \mathrm{d}\pi \, \mathrm{d}P_{\mathrm{ref}}$$
$$\le \frac{C_{\mathrm{adv}} c_s^2}{8\sigma^2}, \tag{97}$$

so that $\Phi_k \ge s - C_{\mathrm{adv}} k c_s^2/(8\sigma^2)$. The bound is applied to the likelihood vector of each history, so it covers adaptive queries and their dependence on the unknown bits; if the algorithm stops before k queries, the history is completed by uninformative symbols.

Decoding from the output. After the k queries, the algorithm outputs a policy $\widehat{P}(\mathsf{h}_k)$. Draw $Y \sim \widehat{P}(\mathsf{h}_k)$ and append it to the history. This conditional sampling rule is common to all ω, so L_k and Φ_k are unchanged.

The word $\varpi(\omega)$ determines $C_{2s-1}(\omega)$, and different words give disjoint cells. Decode $\hat{\varpi}$ as the word whose cell contains Y, choosing an arbitrary word if Y lies in none. Then

$$\{\hat{\varpi} \ne \varpi(\omega)\} \subseteq \{Y \notin C_{2s-1}(\omega)\}.$$

For an augmented history h, define the error set and its Bayes probability by

$$\mathcal{M}_{\mathsf{h}} = \{\omega : \hat{\varpi}(\mathsf{h}) \ne \varpi(\omega)\},$$
$$\epsilon = \int \|v_{\mathsf{h}} \mathbb{1}_{\mathcal{M}_{\mathsf{h}}}\|^2_{L^2(\pi)} \, \mathrm{d}P_{\mathrm{ref}}(\mathsf{h}).$$

If $\mathcal{K}(\omega, \omega') > 0$, the two words differ, so the decoder must err on at least one of them. Symmetry of $\mathcal{K}$, $\|\Gamma\| \le s$, and Cauchy–Schwarz give

$$\Phi_k \le 2 \int \langle v_{\mathsf{h}} \mathbb{1}_{\mathcal{M}_{\mathsf{h}}}, \Gamma v_{\mathsf{h}} \rangle \, \mathrm{d}P_{\mathrm{ref}} \le 2s \left(\int \|v_{\mathsf{h}} \mathbb{1}_{\mathcal{M}_{\mathsf{h}}}\|^2 \, \mathrm{d}P_{\mathrm{ref}} \right)^{1/2} \left(\int \|v_{\mathsf{h}}\|^2 \, \mathrm{d}P_{\mathrm{ref}} \right)^{1/2} = 2s\sqrt{\epsilon}.$$

Combining this with (97) gives

$$\sqrt{\epsilon} \ge \frac{1}{2} \left[1 - \frac{C_{\mathrm{adv}} k c_s^2}{8\sigma^2 s} \right]_+ .$$

Regret. If $Y \notin C_{2s-1}(\omega)$, then $Y \in E_i(\omega)$ for some $i \le 2s - 1$. By (17),

$$G - f_\omega(Y) \ge \kappa T_i \ge \kappa T_{2s-1}.$$

The Bayes regret at $\rho = 0$ is therefore at least

$$\kappa T_{2s-1} \Pr\big(Y \notin C_{2s-1}(\omega)\big) \ge \kappa T_{2s-1} \epsilon,$$

which proves (91).

Finally, the stated choice $s = s_k$ makes

$$\frac{C_{\mathrm{adv}} k c_s^2}{8\sigma^2 s} = \frac{C_{\mathrm{adv}} k}{8\sigma^2 \zeta(\gamma)^2 s^{2\gamma+1}} \le \frac{1}{2}.$$

Hence $B_{\pi_{s_k}}(k, 0) \ge \kappa T_{2s_k - 1}/16 \ge c(2s_k)^{-(\gamma-1)}$ by (16). Since $s_k \le C(1+k)^{1/(2\gamma+1)}$, this proves (92). □

F.4 Small Regularization: Conclusion

For $\rho = 0$, Lemma 34 with $k = \max\{1, \lceil 2t \rceil\}$ and the product prior $\pi_{s_k}^{\otimes M}$ (the primitives of different prompts are independent copies of the two-choice family, and the prompts are exogenous), together with (92) and $1 + k \le 3(1 + t)$, gives

$$R^*(n, M, 0) \ge \frac{1}{2} B_{\pi_{s_k}}(k, 0) \ge c(1+t)^{-p_0}.$$

For $\rho \in (0, 1]$, (90) and the identity

$$\rho^{1-1/\gamma} = (1+t)^{-p_0} u^{p_0}, \qquad u = (1+t)\rho^{2+1/\gamma} \tag{98}$$

(note that $(2 + 1/\gamma)p_0 = 1 - 1/\gamma$) give

$$R^*(n, M, \rho) \ge [c - U u^{p_0}]_+ (1+t)^{-p_0} \ge \frac{c}{2}(1+t)^{-p_0}$$

whenever $u \le c_0$ with $U c_0^{p_0} \le c/2$, which is the first case of (11). □

F.5 Large Budget: The Flat-Band Family

By Lemma 34 and $1+k \le 3(1+t)$, it now suffices to exhibit, for a single prompt and every $k \ge 1$, a prior $\pi_{k,\rho}$ with

$$B_{\pi_{k,\rho}}(k,\rho) \ge c\rho \min\left\{1, \left[(1+k)\rho^{2+1/\gamma}\right]^{-p}\right\}. \tag{99}$$

Let $r_c = 1/8$, $a = d\log(1/q)$ and let $Q_\flat$ be the cube in which every primitive of this family places its maximizer:

$$Q_\flat = [1/4, 1/4+q]^d, \qquad \mathsf{T}(u) = \tfrac{1}{4}\mathbf{1} + qu, \qquad z = \frac{1}{4(1-q)}\mathbf{1}, \qquad w = \tfrac{3}{4}\mathbf{1},$$

so that z is the fixed point of T. Write $e_1 = (1,0,\dots,0) \in \mathbb{R}^d$ for the first coordinate vector. Use the bump ψ_v from (86), and for $v_i = z + t_i e_1$ with $q/16 \le t_i \le q/8$ define the base family

$$g_1 = \psi_{v_1}, \qquad g_i = \psi_{v_i} - \psi_w \quad (i \ge 2). \tag{100}$$

By **(A3)** and $\lambda \ge \Lambda_{\alpha,d} \ge 2 \cdot 8^\beta$,

$$\left(\frac{q}{r_c}\right)^\beta \le \frac{c_{\min}}{16 \cdot 3^\gamma}, \qquad \text{so} \qquad q \le \frac{r_c}{48},$$

and v_i is at distance at least $q/4$ from the boundary of $Q_\flat$. Hence the maximizer of g_i is $v_i \in Q_\flat$, the maximum is 1, the placement cube is $Q_\flat$ at every level, $|g_i^0|_\alpha \le \lambda$, $\int g_i = 0$ for $i \ge 2$, and the ranges are as required. The value at the fixed point,

$$a_i := g_i(z) = \psi_0(t_i e_1) \in [a_-, a_+], \qquad a_- = \psi_0((q/8)e_1), \quad a_+ = \psi_0((q/16)e_1),$$

can be chosen continuously in $[a_-, a_+] \subset (3/4, 1)$, and $g_i(u) \le a_i$ for $u \notin Q_\flat$. Indeed, radial monotonicity gives a strictly positive width $a_+ - a_- > 0$ depending only on the structural parameters, and $1 - a_- \le B_{\alpha,d}(q/8)^\beta < 1/4$ by **(A3)**. For $u \notin Q_\flat$, the distance from u to v_i in either relevant radial norm is at least $q/4 > t_i$; radial monotonicity and $\psi_w \ge 0$ therefore give $g_i(u) \le a_i$.

Flat band. Let

$$\tau = \rho^{-1/\gamma}, \qquad u_0 = \left(\frac{a_-}{a\zeta(\gamma)}\right)^{1/\gamma}, \qquad v_0 = \left(\frac{a_+}{a\zeta(\gamma)}\right)^{1/\gamma}, \qquad \eta_0 = v_0 - u_0 > 0,$$

$\tau_0 = \max\{1, 2/u_0, 4/\eta_0\}$ and $\rho_{\text{flat}} = \tau_0^{-\gamma}$. Assume first $0 < \rho \le \rho_{\text{flat}}$, and let

$$I = \lceil u_0\tau \rceil \ge 2, \qquad J = \lfloor v_0\tau \rfloor, \qquad L = J - I + 1 \ge \frac{\eta_0\tau}{2}.$$

Choose the values at the fixed point as

$$a_i = a_- \ \ (i < I), \qquad a_i = a\rho\zeta(\gamma) i^\gamma \ \ (I \le i \le J), \qquad a_i = a_+ \ \ (i > J), \tag{101}$$

so that $c_i a_i = a\rho$ inside the band; all values lie in $[a_-, a_+]$. With $H_i = \sum_{j \le i} c_j a_j$ and

$$B_i = q^{d(i-1)} e^{H_{i-1}/\rho}, \qquad \frac{B_{i+1}}{B_i} = \exp\left(-a + \frac{c_i a_i}{\rho}\right),$$

B_i is nondecreasing up to I, constant $=: B$ on $I, \dots, J+1$, and nonincreasing afterwards. On the shell $E_i = \mathsf{T}^{i-1}(\Omega \setminus Q_\flat)$, writing $y = \mathsf{T}^{i-1}(u)$,

$$f(y) = H_{i-1} + c_i g_i(u) + e_i(u), \qquad |e_i(u)| \le \lambda \sum_{j<i} c_j q^{\beta(i-j)} \le \epsilon_{\text{anc}}\, c_{i+1}, \qquad \epsilon_{\text{anc}} = \frac{c_{\min}}{4}, \tag{102}$$

by the geometric series in the proof of Lemma 6. With $D_\flat = (1+\epsilon_{\text{anc}})a/a_-$, the shell masses $w_i = \int_{E_i} e^{f/\rho}\, \mathrm{d}\mu$ satisfy

$$(1-q^d)e^{-D_\flat} B \le w_i \le (1-q^d)e^{D_\flat} B \ \text{ inside the band}, \qquad w_i \le e^{2a} B_{i+2} \ \text{ in general},$$

and since $B_{i+1}/B_i \le e^{-a/2}$ for $i \ge \lceil 2v_0\tau \rceil$, the normalizing constant satisfies

$$\frac{1}{2}\eta_0(1-q^d)e^{-D_\flat}\tau B \le Z = \sum_{i \ge 1} w_i \le C_Z \tau B. \tag{103}$$

Thus the Gibbs policy P_ρ^* spreads a mass of order $1/\tau$ over each of the $L \asymp \tau$ shells of the band.

Perturbations. Let $r_U = \frac{1}{32}(3/4)^{1/\alpha}$ and $U = \frac{1}{2}\mathbf{1} + [-r_U, r_U]^d$, a cube disjoint from $Q_\flat$ and from the supports of the positive and negative bumps. Partition U into N^d cells of side $h = 2r_U/N$, and let ϕ_j be the copy on the j-th cell of the balanced bump on the unit cell,

$$\phi = b_{v_+} - b_{v_-}, \qquad b_v(u) = \psi_v(u), \qquad v_\pm = (1/4 \text{ or } 3/4,\, 1/2, \ldots, 1/2);$$

thus $|\phi_j| \le 1$, $\int \phi_j = 0$, $\int \phi_j^2 = h^d v_\phi$ with

$$v_\phi = 2\int_{\mathbb{R}^d} \psi_0(u)^2 \,\mathrm{d}u = \begin{cases} \dfrac{4\alpha^2}{4^d(d+\alpha)(d+2\alpha)}, & 0 < \alpha \le 1, \\ \dfrac{2r_c^d \pi^{d/2}\Gamma(2m_\alpha + 1)}{\Gamma(2m_\alpha + 1 + d/2)}, & \alpha > 1, \end{cases}$$

which is strictly positive. The zero extensions obey $|\phi_j^0|_\alpha \le 2B_{\alpha,d}h^{-\alpha}$ by the scaling and gluing bounds above. Let

$$\kappa_b = \min\left\{1,\ \frac{a}{4a_+},\ \frac{a(2r_U)^\alpha r_c^{-\alpha}}{4\cdot 8^\alpha a_+}\right\}, \qquad \delta_{\mathrm{p}} = \kappa_b \rho N^{-\alpha}.$$

For $I \le i \le J$ and independent signs $\omega_{i,j} \in \{-1, 1\}$, replace g_i by

$$g_i + \frac{\delta_{\mathrm{p}}}{c_i}\sum_{j=1}^{N^d} \omega_{i,j}\phi_j. \tag{104}$$

Since $c_i a_i = a\rho$ on the band, the amplitude $A_i = \delta_{\mathrm{p}}/c_i$ satisfies $A_i \le 1/4$ and $A_i h^{-\alpha} \le \kappa_b a_+/(a(2r_U)^\alpha) \le 1/4$. The original and added bumps have disjoint supports, so the scaling and gluing bounds give $|g_i^0|_\alpha \le 2B_{\alpha,d} \le \lambda$ uniformly in N and the signs. The perturbation has sup norm at most $1/4 < a_-$ and vanishes outside U; it therefore keeps the zero mean, range, unique maximizer v_i, maximum 1, value a_i at z, placement cube $Q_\flat$ and the bound $g_i \le a_i$ off $Q_\flat$. Hence the perturbed instance belongs to $\mathcal{F}_{M,\theta}$ for every sign pattern. The ancestor perturbations vanish on $Q_\flat$, so (102) and (103) hold for all signs, and each physical cell $\mathsf{T}^{i-1}(\mathrm{cell}_j)$ in the band has Gibbs mass

$$b \ge c_{\mathrm{mass}}\frac{h^d}{\tau}, \qquad c_{\mathrm{mass}} = \frac{e^{-D_\flat}}{C_Z}. \tag{105}$$

F.6 Large Budget: Decoding and Information Bound

For laws P, P' with densities p, p' relative to a common law R, use the squared Hellinger distance

$$H^2(p, p') := \int \left(\sqrt{p} - \sqrt{p'}\right)^2 \mathrm{d}R = 2(1 - \mathrm{Aff}(P, P')).$$

Separation of the two hypotheses on a cell. Let w be the conditional density of P_ρ^* on a cell of the band under the unperturbed family; it is within a factor $e^{\pm 2D_\flat}$ of the uniform density on the cell and does not depend on the signs (the ancestor perturbations vanish on $Q_\flat$ and the deeper levels vanish on the shell). With $t_\delta = \delta_{\mathrm{p}}/\rho \le 1$, the conditional densities under the two signs of bit (i, j) are

$$q_\pm = \frac{we^{\pm t_\delta \phi_j}}{\mathbb{E}_w e^{\pm t_\delta \phi_j}},$$

and with $V, V' \sim w$ independent,

$$A_t := \left(\mathbb{E}_w e^{t_\delta \phi_j}\right)\left(\mathbb{E}_w e^{-t_\delta \phi_j}\right) = \mathbb{E}\cosh\left(t_\delta(\phi_j(V) - \phi_j(V'))\right) \in \left[1 + t_\delta^2 \operatorname{Var}_w(\phi_j),\, e^2\right].$$

Hence the squared Hellinger distance satisfies

$$H^2(q_+, q_-) = 2\left(1 - A_t^{-1/2}\right) \ge e^{-3}t_\delta^2 \operatorname{Var}_w(\phi_j) \ge c_{\mathrm{sep}} t_\delta^2, \qquad c_{\mathrm{sep}} = e^{-3-2D_\flat} v_\phi. \tag{106}$$

Decoding. Given any output policy with density $\hat{p}$, decode the sign of bit (i,j) by comparing $\langle\sqrt{\hat{p}},\sqrt{q_+}\rangle$ and $\langle\sqrt{\hat{p}},\sqrt{q_-}\rangle$ on the physical cell. If the decoded sign is wrong, the squared L^2 distance between $\sqrt{\hat{p}}$ and $\sqrt{b\,q_{\mathrm{true}}}$ on the cell is at least $\frac{b}{4}H^2(q_+,q_-)$ (the distance to the half-space of the wrong decision). Using $\mathrm{KL}(P\parallel Q')\geq\int(\sqrt{p}-\sqrt{q'})^2$, $\mathcal{R}_\rho(P)=\rho\,\mathrm{KL}\left(P\,\middle\|\,P^*_\rho\right)$, (105) and (106), and summing over cells,

$$\mathbb{E}\,\mathcal{R}_\rho(\widehat{P})\geq\frac{c_{\mathrm{mass}}c_{\mathrm{sep}}\delta_{\mathrm{p}}^2h^d}{4\rho\tau}\sum_{i=I}^{J}\sum_{j=1}^{N^d}\Pr\left(\text{bit }(i,j)\text{ decoded wrongly}\right),\tag{107}$$

where the probability is over the prior and the transcript. Neither the unknown normalizing constant nor the mass that $\widehat{P}$ puts outside the cells affects this bound; if $\widehat{P}$ is singular, the regret is infinite.

Information. Let $\Phi_{i,j}=\phi_j\circ\mathsf{T}^{-(i-1)}$ on $\mathsf{T}^{i-1}(U)$ and 0 elsewhere. Flipping the sign (i,j) changes the reward by $2\delta_{\mathrm{p}}\Phi_{i,j}$, and since the $\Phi_{i,j}$ have disjoint supports and $|\Phi_{i,j}|\leq 1$, Jensen's inequality gives, for any cube D,

$$\sum_{i,j}\langle\Phi_{i,j}\rangle_D^2\leq\left\langle\sum_{i,j}\Phi_{i,j}^2\right\rangle_D\leq 1,\tag{108}$$

where $\langle\cdot\rangle_D$ is the average over D. By Lemma 35, for every sign pattern ω,

$$\sum_{i,j}\mathrm{KL}\left(\bar{P}_\omega\,\middle\|\,\bar{P}_{\omega^{(i,j)}}\right)\leq\frac{2k\delta_{\mathrm{p}}^2}{\sigma^2}.\tag{109}$$

By the two-point test and Pinsker's inequality (Tsybakov, 2008, proof of Theorem 2.2(i) and Lemma 2.5(i)) averaged over all LN^d bits and all sign patterns, and Jensen's inequality on the square root,

$$2^{-LN^d}\sum_{\omega}\sum_{i,j}\Pr_{\omega}\left(\text{bit }(i,j)\text{ wrong}\right)\geq\frac{LN^d}{2}\left[1-\sqrt{\frac{k\delta_{\mathrm{p}}^2}{LN^d\sigma^2}}\right].\tag{110}$$

Conclusion for small ρ. Choosing

$$N=\max\left\{1,\left\lceil\left(\frac{4k\kappa_b^2\rho^2}{L\sigma^2}\right)^{1/(2\alpha+d)}\right\rceil\right\}$$

makes the bracket in (110) at least $1/2$, so the expected number of wrong bits is at least $LN^d/4$. Substituting into (107) with $N^dh^d=|U|$ and $L/\tau\geq\eta_0/2$,

$$B_{\pi_{k,\rho}}(k,\rho)\geq C_0'\kappa_b^2\rho N^{-2\alpha},\qquad C_0'=\frac{c_{\mathrm{mass}}c_{\mathrm{sep}}\eta_0|U|}{32},$$

and since $N^{2\alpha+d}\lesssim 1+k\rho^{2+1/\gamma}$ (recall $L\asymp\tau=\rho^{-1/\gamma}$), we obtain (99) for $0<\rho\leq\rho_{\mathrm{flat}}$ with a constant C_{small} accounting for the rounding.

Large ρ. For $\rho_{\mathrm{flat}}\leq\rho\leq 1$, use the base family with all $a_i=a_-$ and perturb only level 2, with amplitude $b_0N^{-\alpha}$, where

$$b_0=\min\left\{\frac{1}{4},\ \frac{r_c^{-\alpha}(2r_U)^\alpha}{4\cdot 8^\alpha},\ \frac{\rho_{\mathrm{flat}}}{c_2}\right\},\qquad\delta_{\mathrm{p},0}=c_2b_0.$$

The inclusion in $\mathcal{F}_{M,\theta}$ is verified as before, since $b_0N^{-\alpha}h^{-\alpha}=b_0/(2r_U)^\alpha\leq 1/4$; $t_\delta=\delta_{\mathrm{p},0}N^{-\alpha}/\rho\leq 1$, the true mass of a cell is at least $q^de^{-2/\rho_{\mathrm{flat}}}h^d$ since $|f|\leq 1$, and the conditional density is within a factor $e^{\pm 2/\rho_{\mathrm{flat}}}$ of uniform. The same decoding and information bounds with $N=\max\{1,\lceil(4k\delta_{\mathrm{p},0}^2/\sigma^2)^{1/(2\alpha+d)}\rceil\}$ give

$$B_{\pi_{k,\rho}}(k,\rho)\geq C_{\mathrm{large}}\min\{1,k^{-p}\},$$

and since $\rho\min\{1,(k\rho^{2+1/\gamma})^{-p}\}\leq\max\{1,\rho_{\mathrm{flat}}^{1-(2+1/\gamma)p}\}\min\{1,k^{-p}\}$ on this range, (99) holds with

$$c=\min\left\{C_{\mathrm{small}},\ \frac{C_{\mathrm{large}}}{\max(1,\rho_{\mathrm{flat}}^{1-(2+1/\gamma)p})}\right\};$$

ρ_{flat} depends only on the structural parameters.

F.7 Large Budget: Conclusion

Combining (99) with Lemma 34 and $1 + k \le 3(1 + t)$ yields (11). The second statement of the large-budget bound of Theorem 4 follows since $n \ge M\rho^{-(2+1/\gamma)}$ implies $(n + M)\rho^{2+1/\gamma}/M \le 2n\rho^{2+1/\gamma}/M$. □

G Proofs for Section 5

Recall the protocol of $\mathcal{A}_{\mathrm{unif}}$: at round r', $X_{r'} \sim \nu$ and $Y_{r'} \sim \mu$ are generated independently of the past; the learner chooses a depth $H_{r'} \in \mathbb{N}_0$ from the history, $(X_{r'}, Y_{r'})$ and internal randomness, queries $D_{r'} = D_{H_{r'}}(Y_{r'})$ and observes $Z_{r'} = Q(X_{r'}, D_{r'}) + \xi_{r'}$; after n rounds it outputs an arbitrary policy. We write

$$t = \frac{n}{M}, \qquad g = \gamma - 1, \qquad s = 2 + \frac{1}{\gamma}, \qquad L(t) = \log(e + t).$$

The two bounds of Theorem 5 are proved in Appendix G.1 and Appendix G.3, and Appendices G.2 and G.4 show that neither is an artefact of the method of proof: the first bound is attained by a fixed query depth, and for a fixed ρ the exponent of n is not worsened by fixed sampling. The two bounds use different families, but for the same reason: a query made from outside the cell that carries the unknown information cannot see it, whatever its depth, and under uniform sampling that cell is reached only exponentially rarely. The logarithmic bound uses the two-choice family of Appendix F.2, which lies in $\mathcal{F}_{M,\theta}$ whenever $\lambda \ge \Lambda_{\alpha,d}$, and the following form of that property; the exponential bound reuses the flat-band family of Appendix F.5 and applies the same volume factor to the information bound of Appendix F.6.

Lemma 38 (Cancellation)**.** *Fix $k \ge 2$, set all bits to 0 except $\omega_{k,m} = b_m$, and let f_0, f_1 be the two rewards of prompt m (with $b_m = 0, 1$). Then, with $C_{k-1} = C_{k-1}(m)$ common to both and $\mu(C_{k-1}) = v_k := 2^{-d\ell_\star(k-1)}$,*

$$f_0 = f_1 \ \text{ on } \Omega \setminus C_{k-1}, \qquad \int_{C_{k-1}} (f_0 - f_1)\,\mathrm{d}\mu = 0, \qquad \int_D (f_0 - f_1)\,\mathrm{d}\mu = 0 \ \text{ if } D \not\subset C_{k-1}. \tag{111}$$

In particular, the two oracle cube averages agree whenever $D \not\subset C_{k-1}$.

Proof. The first $k - 1$ levels are common, all deeper levels are supported in C_{k-1}, and each has zero integral by **(A4)**. A dyadic cube that is not contained in C_{k-1} is either disjoint from C_{k-1} or contains it; in both cases the cube averages coincide. □

G.1 Proof of the Logarithmic Bound of Theorem 5

Fix $k \ge 2$ and use the prior of Lemma 38 with independent fair bits b_m. Let

$$E_m = \bigcap_{r' \le n} \{X_{r'} \ne m \ \text{ or } \ Y_{r'} \notin C_{k-1}\};$$

by the exogenous laws,

$$\Pr(E_m) = \Big(1 - \frac{v_k}{M}\Big)^n \tag{112}$$

exactly. Couple the two experiments $b_m = 0, 1$ by using the same other bits, internal randomness, $(X_{r'}, Y_{r'})$ and noise. On E_m, by induction over rounds, the histories coincide: the same depth is chosen from the same history, and the observation is the same by Lemma 38 when $X_{r'} = m$ (the queried cube contains $Y_{r'} \notin C_{k-1}$) and trivially otherwise. Hence, conditionally on E_m, the bit b_m is fair and independent of the transcript and of the output policy; this holds also for $\sigma = 0$.

Let $G_b = \max f_b$. By (17), $G_b - f_b(y) \ge \kappa T_k$ for $y \notin C_k^b$, and C_k^0, C_k^1 are disjoint, so for every y at least one of the two losses is at least κT_k. Integrating the common output policy and averaging the fair bit,

$$\mathbb{E}\Big[G_{b_m} - \mathbb{E}_{\widehat{P}(\cdot|m)} f_{b_m} \ \Big|\ E_m\Big] \ge \frac{\kappa T_k}{2}.$$

Since the unregularized regret is nonnegative on E_m^c, averaging over prompts gives, for all $n, M \geq 1$ and $k \geq 2$,

$$R^*_{\mathrm{unif}}(n, M, 0) \geq \frac{\kappa T_k}{2}\left(1 - \frac{2^{-d\ell_\star(k-1)}}{M}\right)^n. \tag{113}$$

For $\rho > 0$, Lemma 36 gives $\mathcal{R}_\rho(P) \geq \mathcal{R}_0(P) - U\rho^{g/\gamma}$ for every policy, with the same constant U (singular policies have infinite regularized regret). Therefore, for $0 \leq \rho \leq 1$,

$$R^*_{\mathrm{unif}}(n, M, \rho) \geq \left[\frac{\kappa T_k}{2}\left(1 - \frac{2^{-d\ell_\star(k-1)}}{M}\right)^n - U\rho^{g/\gamma}\right]_+. \tag{114}$$

Choose

$$k = 2 + \left\lceil \frac{\log(2(1+t))}{d\ell_\star \log 2} \right\rceil,$$

so that $tv_k \leq 1/2$ and, by Bernoulli's inequality, $(1 - v_k/M)^n \geq 1 - nv_k/M \geq 1/2$. Since $k \leq CL(t)$ and $T_k \geq c_T k^{-g}$, we get $R^*_{\mathrm{unif}} \geq [c_{\mathrm{L}} L(t)^{-g} - U\rho^{g/\gamma}]_+$, and the second claim of the first case of (12) follows with a such that $Ua^{g/\gamma} \leq c_{\mathrm{L}}/2$. □

G.2 A Uniform-Sampling Method Attaining the Logarithmic Rate

Proposition 39. *Let $H = \lfloor \log_2(1+t)/(2d) \rfloor$ and $B_H = 2^{dH}$. The algorithm in $\mathcal{A}_{\mathrm{unif}}$ that queries $D_H(Y_{r'})$ at every round, averages the observations of each (m, D) (clipped to $[-1, 1]$, 0 if unobserved) and outputs the uniform distribution on the cube with the largest average satisfies, for all $n, M \geq 1$, $\rho \in [0, 1]$ and $\sigma \geq 0$,*

$$R^*_{\mathrm{unif}}(n, M, \rho) \leq C\big[L(t)^{-g} + \rho L(t)\big]. \tag{115}$$

Consequently, $R^*_{\mathrm{unif}}(n, M, \rho) = \Theta(L(t)^{-g})$ whenever $\lambda \geq \Lambda_{\alpha,d}$ and $\rho \leq aL(t)^{-\gamma}$.

Proof. For each (m, D) the expected number of observations is

$$\lambda_H = \frac{t}{B_H} \geq \frac{t}{\sqrt{1+t}} \geq \sqrt{\frac{2t}{3}} \qquad (t \geq 2).$$

By the exponential supermartingale $\exp\big(\eta' S - \sigma^2 \eta'^2 N/2\big)$ of the noise sums and the lower binomial tail of the counts, for $0 < x \leq 2$,

$$\Pr\left(\left|\widehat{Q}(m, D) - Q(m, D)\right| > x\right) \leq e^{-\lambda_H/8} + 2\exp\left(-\frac{\lambda_H x^2}{4\sigma^2}\right).$$

A union bound over the B_H cubes of one prompt and integration of the tail give, for $t \geq 2$,

$$\mathbb{E}\max_D \left|\widehat{Q}(m, D) - Q(m, D)\right| \leq 2B_H e^{-\lambda_H/8} + C\sigma\sqrt{\frac{\log(eB_H)}{\lambda_H}} \leq C_\sigma t^{-1/4}\sqrt{L(t)}.$$

Let D^*_H be the cube of depth H containing the maximizer and $K' = \lfloor H/\bar{\ell} \rfloor$; then $D^*_H \subset C_{K'}$ and $G_m - Q(m, D^*_H) \leq C(H+1)^{-g}$ by (18). The empirically best cube $\widehat{D}_m$ therefore satisfies

$$G_m - Q(m, \widehat{D}_m) \leq C(H+1)^{-g} + 2\max_D \left|\widehat{Q}(m, D) - Q(m, D)\right|,$$

and the KL divergence of the output is exactly $\log B_H = dH \log 2$. Averaging over prompts (no simultaneous success is needed) and treating small t by the uniform output gives (115); the matching lower bound is Theorem 5. □

G.3 Proof of the Exponential Bound of Theorem 5

Fix $\sigma > 0$. We reuse the flat-band family of Appendix F.5, now with independent signs $\omega_{i,m,j}$ for each prompt m and with the common placement cube $Q_\flat$ at all levels (which may be revealed to the learner). The decoding bound (107) holds for each prompt, with an additional factor $1/M$ from the prompt average:

$$\mathbb{E}\,\mathcal{R}_\rho(\widehat{P}) \geq c\frac{\delta_{\mathrm{p}}^2 h^d}{\rho\tau M}\sum_{i=I}^{J}\sum_{m=1}^{M}\sum_{j=1}^{N^d}\Pr\left(\text{sign } (i, m, j) \text{ wrong}\right). \tag{116}$$

Information under uniform sampling. All perturbations are supported in the cube $C_{I-1} = \mathsf{T}^{I-1}(\Omega)$, of volume $v_I = q^{d(I-1)}$, and have zero mean. If a uniform response $Y_{r'} \notin C_{I-1}$, then any dyadic cube $D \ni Y_{r'}$ is disjoint from C_{I-1} or contains it, and in both cases $\langle \Phi_{i,j} \rangle_D = 0$ for all (i, j). If $Y_{r'} \in C_{I-1}$, then $\sum_{i,j} \langle \Phi_{i,j} \rangle_D^2 \le 1$ by (108). Hence, conditionally on the history and $X_{r'}$, the expected squared sensitivity of the r'-th observation to the signs is at most v_I, whatever depth is chosen from $(X_{r'}, Y_{r'})$ and the history. By Lemma 35 and the chain rule, for every sign pattern,

$$\sum_{i,m,j} \mathrm{KL}\left(\bar{P}_\omega \,\middle\|\, \bar{P}_{\omega^{(i,m,j)}}\right) \le \frac{2n\delta_{\mathrm{p}}^2 v_I}{\sigma^2}, \tag{117}$$

where only the signs of the arriving prompt contribute at each round, so no extra factor M appears.

Conclusion. With $B_{\mathrm{bits}} = LMN^d$ signs, the two-point test and Pinsker's inequality give an expected number of wrong signs at least

$$\frac{B_{\mathrm{bits}}}{2}\left[1 - \sqrt{\frac{n\delta_{\mathrm{p}}^2 v_I}{B_{\mathrm{bits}}\sigma^2}}\right],$$

and the choice

$$N = \max\left\{1, \left\lceil \left(\frac{4n\kappa_b^2\rho^2 v_I}{LM\sigma^2}\right)^{1/(2\alpha+d)} \right\rceil\right\}$$

makes it at least $B_{\mathrm{bits}}/4$. Substituting into (116) with $N^d h^d = |U|$ and $L \asymp \tau$,

$$R^*_{\mathrm{unif}}(n, M, \rho) \ge c\rho N^{-2\alpha} \ge c'\rho \min\left\{1, \left[\frac{1}{t\rho^{2+1/\gamma} v_I}\right]^p\right\}.$$

For $\rho \le \rho_0$ small enough, $I - 1 \ge u_0\tau/2$, so that

$$v_I \le \exp\left(-c_0\rho^{-1/\gamma}\right), \qquad c_0 = \frac{au_0}{2},$$

which gives the second case of (12); the second statement follows by selecting the second term of the minimum. □

G.4 Fixed Regularization: the Exponent in n Is Not Worsened

Proposition 40. *For all $n, M \ge 1$, $\rho \in (0, 1]$ and $\sigma \ge 0$, there is an algorithm in $\mathcal{A}_{\mathrm{unif}}$ with*

$$R^*_{\mathrm{unif}}(n, M, \rho) \le \min\left\{2,\ C\frac{e^{2/\rho}}{\rho}\left[\frac{L(t)^2}{1+t}\right]^p\right\}. \tag{118}$$

Consequently, for $\sigma > 0$, $\lambda \ge \Lambda_{\alpha,d}$ and $t \ge C_0\rho^{-s}$,

$$c\rho^{1-sp}(1+t)^{-p} \le R^*_{\mathrm{unif}}(n, M, \rho) \le C\frac{e^{2/\rho}}{\rho}\left[\frac{L(t)^2}{1+t}\right]^p,$$

so that, for each fixed $\rho > 0$, the optimal exponent of t is p.

Proof. Finite approximation family. For $0 < \varepsilon \le 1/4$ let $K' = \lceil \frac{2}{d\ell_\star} \log_2(1/\varepsilon) \rceil$, so that

$$\left\| f - f^{[K']} \right\|_{L^2(\mu)} \le \mu(C_{K'})^{1/2} \le \varepsilon.$$

On each cell of a dyadic mesh of side $h_T \asymp \varepsilon^{1/\alpha}$, use the degree-$k$ Taylor polynomial of a primitive at the lower-left corner, where $k = \lceil \alpha \rceil - 1$. Taylor's formula and **(A1)** give remainder at most $C_{\alpha,d}\lambda h_T^\alpha$, also for integer α; choose the mesh constant so this is at most $\varepsilon/4$. Quantize the finitely many coefficients on each cell to multiples of $c_{\alpha,d}\varepsilon$; they are bounded by $1 + \lambda$, and quantization adds at most $\varepsilon/4$ on that cell. Clip the resulting piecewise polynomial to $[-1, 1]$. This gives a finite sup-norm $\varepsilon/2$ net with $O(\varepsilon^{-d/\alpha})$ cells and $O_{\alpha,d}(1)$ coefficients per cell. Next let J' be the smallest

nonnegative integer with $\lambda 2^{-\beta J'} \le \varepsilon/2$, and replace every candidate by its averages on the depth-J' grid. Averaging contracts the approximation error, while the true primitive differs from its own grid averages by at most $\lambda 2^{-\beta J'}$. Thus these piecewise constant candidates have sup error at most ε and the same cardinality, with

$$\log N_{\mathrm{prim}} \le C\varepsilon^{-d/\alpha}\log(1/\varepsilon)$$

candidates. Enumerate also all placements of the first K' levels ($O(K')$ bits). Let $\mathcal{G}_\varepsilon$ be the finite set of scores $\sum_{i\le K'} c_i\tilde{h}_i$ obtained from these candidates; then

$$\log|\mathcal{G}_\varepsilon| \le C\varepsilon^{-d/\alpha}\log^2(1/\varepsilon), \qquad \inf_{v\in\mathcal{G}_\varepsilon}\|f-v\|_{L^2(\mu)} \le 2\varepsilon. \tag{119}$$

Least squares on cube averages. Fix the query depth $H = \bar{\ell}K' + J'$; every $v \in \mathcal{G}_\varepsilon$ is constant on the depth-H grid, and the observation is $Z_{r'} = f_H(D_H(Y_{r'})) + \xi_{r'}$, where f_H is the projection of f onto the grid. For each prompt use its first $q' = \lfloor t/2 \rfloor$ observations and let $\hat{v}$ minimize the empirical squared loss over $\mathcal{G}_\varepsilon$. The ERM basic inequality, the conditional sub-Gaussian bound for the noise with a union bound over $\mathcal{G}_\varepsilon$, Bernstein's inequality for the i.i.d. uniform cells and the orthogonality $\|f-f_H\|^2 + \|f_H - v_0\|^2 = \|f-v_0\|^2$ give

$$\mathbb{E}\|\hat{v}-f\|^2_{L^2(\mu)} \le C\left[\varepsilon^2 + \frac{(1+\sigma^2)\varepsilon^{-d/\alpha}\log^2(1/\varepsilon)}{q'}\right]. \tag{120}$$

Choosing $\varepsilon \asymp [\log^2(e+q')/(1+q')]^{\alpha/(2\alpha+d)}$ gives

$$\mathbb{E}\|\hat{v}-f\|^2_{L^2(\mu)} \le C\left[\frac{\log^2(e+q')}{1+q'}\right]^p.$$

Gibbs output. Output $P_{\hat{v}} = P_{\hat{v},\rho}$. Since $f, \hat{v} \in [-1,1]$, the Gibbs densities along the segment between them are bounded by $e^{2/\rho}$, and (72) gives

$$\mathcal{R}_\rho(P_{\hat{v}}) \le \frac{e^{2/\rho}}{2\rho}\|f-\hat{v}\|^2_{L^2(\mu)}.$$

Prompts. If $N_m < q'$, output the uniform distribution (regret at most 2); $\Pr(N_m < q') \le e^{-t/8}$ for $t \ge 4$, and conditionally on the prompt sequence the first q' inputs of a prompt are i.i.d. uniform. Averaging over prompts gives (118). The lower bound in the second statement is the large-budget bound of Theorem 4, which applies to $\mathcal{A}_{\mathrm{unif}}$ as a subclass of all algorithms. □