THEORETICAL FOUNDATIONS OF $\max@k$ REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning is a cornerstone technique for modern large reasoning models. Usually, for difficult tasks such as code generation and theorem proving, the agent is evaluated by generating K responses rather than sampling a single response, and performance is then measured using a retry-aware metric such as $\max@k$. Despite their practical importance, the theoretical foundations of learning under such criteria remain limited. In this work, we provide a theoretical study of the $\max@k$ learning problem in finite-horizon reinforcement learning. We show that optimizing the $\max@k$ objectives is fundamentally different from standard expected-return maximization. In particular, we prove that Markovian policies are in general insufficient, identify a compact state augmentation that restores optimality, and explicitly characterize the performance gap that can arise between history-dependent and non-history-dependent policies. Moreover, we show that learning $\max@k$-optimal policies is statistically harder than standard reinforcement learning and provide an efficient algorithm that achieves the optimal sample complexity rate.

1 INTRODUCTION

Reinforcement Learning is one of the main drivers of the recent successes of Large Language Models (LLMs) in complex coding and mathematical tasks (see, *e.g.*, (Jaech et al., 2024; Lambert et al., 2024; Xin et al., 2024; Guo et al., 2025)). When dealing with challenging tasks such as theorem proving and code generation, the LLM performance can improve substantially if the model is allowed to generate several responses and is evaluated on the best one. Early evidence of this behavior appeared in code generation and mathematics (Chen et al., 2021; Cobbe et al., 2021; Li et al., 2022; Xin et al., 2024), and subsequent studies showed inference-compute scaling laws across a broad range of tasks and models (see, *e.g.*, (Wang et al., 2022; Brown et al., 2024; Snell et al., 2025)). This metric, which scores the model by the highest reward obtained across K generations, is called $\max@k$ and extends $\mathrm{pass}@k$ beyond binary outcomes (Bagirov et al., 2025).

A growing body of literature has shown that optimizing for the classic single-sample objective and then evaluating with the $\max@k$ metric has suboptimal performance (Cui et al., 2025; Chen et al., 2026b). A natural solution to this mismatch is to optimize the $\max@k$ objective directly. Recent work has therefore proposed gradient estimators for $\max@k$, sometimes referred to as inference-aware estimators, as well as modifications of advantage estimates to better align policy updates with the objective (Chow et al., 2025; Chen et al., 2025; Tang et al., 2025; Walder & Karkhanis, 2026; Bagirov et al., 2025; Takashiro et al., 2026). While these approaches correctly identify the relevant optimization problem, we still lack a complete theoretical understanding of the $\max@k$ objective in reinforcement learning. Indeed, most existing RL theory relies, explicitly or implicitly, on the linearity of expected cumulative reward, whereas the $\max@k$ objective is nonlinear in the distribution over trajectories.

In this paper, we take on the challenging task of analyzing the finite-horizon model under the $\max@k$ objective from a theoretical perspective, clearly highlighting the differences between classic RL and $\max@k$ RL. Our main contributions can be summarized as follows.

1. In Section 3 we show that Markovian policies are not optimal for this problem. This is a pivotal difference between standard RL and RL with $\max@k$ objectives. We also prove

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.

that the gap between these two classes of policies can be large, and that the ratio between them is lower-bounded by a constant greater than 1. This shows that considering history-dependent policies yields a material improvement in performance over Markovian policies.

2. In Section 3.1, we show that full history dependency is not necessary. Indeed, we demonstrate that for optimal policies, the history can be compressed into only two relevant pieces of information: the maximum reward accrued in past rollouts $\{1, \dots, j-1\}$, and the cumulative reward inside the current j-th rollout. This defines a class of policies that is much more manageable than full history-dependent policies, making working with history-dependent policies not computationally much more demanding than standard RL.
3. In Section 3.2, we exploit our compressed history representation for the offline planning problem and show that we can efficiently approximate optimal history-dependent policies. In particular, we demonstrate that computing an ϵ-optimal policy can be done in time polynomial in $1/\epsilon, K, S, A, H$. This shows that the computational problem of finding an optimal policy remains tractable even if optimal policies cannot be oblivious to history.
4. In Section 3.3, we show that *approximating* the optimal history-dependent policy is necessary, and that inverse polynomial dependence on the approximation is necessary, since even for inverse exponential approximations the problem becomes NP-hard. This contrasts with standard finite-horizon MDPs, where the *exact* planning problem can be solved in polynomial time.
5. In Section 4, we turn to the learning problem, where we first show that, compared to the classic bound for learning under a generative model in a finite-horizon MDP (*e.g.*, Gheshlaghi Azar et al., 2013; Sidford et al., 2019), the sample complexity suffers an additional multiplicative linear dependence on the number of rollouts K. This follows from the fact that small probabilities of reaching a large payoff must be estimated with much greater precision, since even small probabilities can yield a large payoff when sampled multiple times. Formally, we prove a $\Omega(KSAH^3\epsilon^{-2}\log(\delta^{-1}))$ lower bound for any (ϵ, δ)-correct algorithm. We also show a matching $\widetilde{O}(KSAH^3\epsilon^{-2}\log(\delta^{-1}))$ upper bound for the time homogeneous case and a $\widetilde{O}(KSAH^4\epsilon^{-2}\log(\delta^{-1}))$ in the inhomogeneous case. In general, this shows that learning for $\max@k$ is K times more demanding than classical RL.

We now present some high-level connections between our theoretical contributions and some state of the art approaches to reasoning tasks (additional related works are discussed in Appendix A) A recurring recent trend in this literature is inference-time refinement, where multiple candidate solutions are generated sequentially, with later attempts leveraging information from earlier ones. Examples include self-refinement and reflection methods (Madaan et al., 2023; Shinn et al., 2023; Dhuliawala et al., 2024; Zhang et al., 2025a; Han et al., 2026; Gao et al., 2026), search-based reasoning methods (Yao et al., 2023; Besta et al., 2024), recursive self-aggregation (Venkatraman et al., 2025), and multi-attempt reinforcement learning with verifier feedback (Chung et al., 2025; Ildiz et al., 2026). Recent proof-oriented systems also use population-level generation, verification, refinement, and selection at test time (Chen et al., 2026a; Tsoukalas et al., 2026; Chung et al., 2026). Our theoretical model shows that independent sampling of several responses is quantitatively suboptimal with respect to these recent adaptive approaches. At a conceptual level, these methods can be interpreted as a richer implementation of the idea that history-dependent policy over reasoning traces yields larger performance. At the same time, our compression result shows that the full history is not necessary: for optimal control, it suffices to retain the best return from previous rollouts and the cumulative return from the current rollout. This mirrors, at an abstract level, the practical design principle behind many adaptive reasoning methods: useful information from previous attempts should be retained, but only through a compressed summary rather than by conditioning on the entire history. Overall, our theoretical model not only explores an interesting new objective function for RL motivated by the current ways reasoning models are deployed and evaluated, but also aligns with the evidence and principles developed by practitioners facing these problems.

2 Preliminaries

Markov Decision Process. A finite horizon *time inhomogeneous* Markov Decision Process (MDP, Puterman, 1990) is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h=1}^H, \{r_h\}_{h=1}^H)$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$

is a finite set of actions, $H \in \mathbb{N}$ is the interaction horizon, $\{p_h\}_{h=1}^H$ is the transition kernel with $p_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$[1] assigning to each state-action pair (s, a) a probability distribution $p_h(\cdot \mid s, a)$ over the next states, and $\{r_h\}_{h=1}^H$ is the the reward function with $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ assigning the immediate reward obtained after taking action a in state s at step h. The MDP $\mathcal{M}$, instead, is said to be *time homogeneous* if the transition kernel and reward function are fixed across the interaction horizon, *i.e.*, all p_h's and r_h's are equal to some fixed p and r. A history-dependent stochastic policy π models the behavior of the agent and is defined as a sequence $\pi = \{\pi_h\}_{h\in[H]}$, where each π_h maps the history $\mathcal{H}_h$ observed up to the beginning of step h to $\Delta(\mathcal{A})$. Precisely, the history $\mathcal{H}_h$ is an element that belongs to $\mathcal{S} \times (\mathcal{S} \times \mathcal{A})^{h-1}$, and the interaction with the environment proceeds as follows. The agent starts in some state $s_1 \in \mathcal{S}$, and for each $h \in [H]$, it selects an action $a_h \sim \pi_h(\cdot \mid \mathcal{H}_h)$, receives a reward $r_h(s_h, a_h)$ and transitions to the next state $s_{h+1} \sim p_h(\cdot \mid s_h, a_h)$. The interaction with the environment ends after H actions. Given an initial state $s \in \mathcal{S}$, the objective of the agent is to maximize the expected cumulative reward over the horizon $\mathbb{E}_\pi \left[\sum_{h\in[H]} r_h(s_h, a_h) \mid s_1 = s\right]$, where the expectation is taken with respect to the stochasticity of the environment and the policy. A particularly important class of policies is the Markovian one, for which the action distribution at step h depends only on the current state, *i.e.*, $\pi_h(\cdot \mid s_h)$. It is well known that in finite-horizon MDPs, Markovian policies are sufficient for optimality: there always exists a Markovian policy that maximizes the expected cumulative reward.

The $\max@k$ Decision Process. In this work, we extend the above formulation by considering an agent with a budget of $K \in \mathbb{N}$ rollouts. For a starting state $s \in \mathcal{S}$, the agent interacts with the MDP for $K \in \mathbb{N}$ rollouts, each of length H. At the beginning of every rollout, the environment is reset to the initial state s, while the agent retains the complete history collected during all previous rollouts. Formally, we denote by $\mathcal{H}_{h,i} \in (\mathcal{S} \times (\mathcal{S} \times \mathcal{A})^H)^{i-1} \times (\mathcal{S} \times (\mathcal{S} \times \mathcal{A})^{h-1})$ the *entire* history observed up to the beginning of step h in rollout i. Then, the behavior of the agent is modeled by history-dependent stochastic policy $\pi = \{\pi_{h,i}\}_{h\in[H],i\in[K]}$, where each $\pi_{h,i}$ maps the history $\mathcal{H}_{h,i}$ to a distribution over actions. In this document, we denote by Π_{hist} the set of all the history dependent policies. For an initial state $s \in \mathcal{S}$, the performance of any policy $\pi \in \Pi_{\text{hist}}$ is no longer the expected return of a single trajectory, but the *expected best return* observed among the K rollouts, *i.e.*,

$$\mathbb{E}_\pi \left[\max_{i\in[K]} \sum_{h=1}^{H} r_h(s_{h,i}, a_{h,i}) \mid s_{1,i} = s \; \forall i \in [K]\right], \tag{1}$$

where the expectation is taken with respect to the joint distribution over the K rollouts induced by π and the stochasticity of the environment, and a state-action pair $(s_{h,i}, a_{h,i})$ denotes the state-action pair of the i-th rollout at step h. We first observe that any history dependent policy π satisfies the following Bellman recursive equations. Precisely, for a fixed history-dependent policy π, step $h \in [H + 1]$ and rollout index $i \in [K]$, and history $\mathcal{H}_{h,i}$, define

$$V^\pi_{h,i}(\mathcal{H}_{h,i}) = \mathbb{E}_\pi \left[\max_{j\in[K]} \sum_{l=1}^{H} r_l(s_{l,j}, a_{l,j}) \mid \mathcal{H}_{h,i}\right],$$

as the value function conditioned on the history up to the beginning of step h in rollout i. Then, an application of the tower law of expectation yields

$$V^\pi_{h,i}(\mathcal{H}_{h,i}) = \sum_{a\in\mathcal{A}} \pi_{h,i}(a \mid \mathcal{H}_{h,i}) \sum_{s'\in\mathcal{S}} p_h(s' \mid s_{h,i}, a) V^\pi_{h+1,i}\left(\mathcal{H}_{h,i} \circ (a, s')\right),$$

where $\mathcal{H}_{h,i} \circ (a, s')$ denotes the history obtained from $\mathcal{H}_{h,i}$ by appending the state-action pair (a, s') to the history. Now, observe that, at the end of rollout $i < K$, the recursion moves to the next rollout, meaning that $V^\pi_{H+1,i}(\mathcal{H}_{H+1,i}) = V^\pi_{1,i+1}(\mathcal{H}_{1,i+1})$. At the end of the last rollout, we have that $V^\pi_{H+1,K}(\mathcal{H}_{H+1,K}) = \max_{j\in[K]} \sum_{h=1}^H r_h(s_{h,j}, a_{h,j})$. Finally, at the beginning, we have that $V^\pi_{1,1}(s)$ is the expected value of the best return gathered in K rollouts following π starting at s.

The optimal value function is defined as the supremum over all the history-dependent policies, *i.e.*, $V^\star_{h,i}(\mathcal{H}_{h,i}) = \sup_{\pi\in\Pi_{\text{hist}}} V^\pi_{h,i}(\mathcal{H}_{h,i})$. Via backward induction one can easily show that it satisfies the

[1]Given a set $\mathcal{X}$, we denote by $\Delta(\mathcal{X})$ the set of probability distributions over $\mathcal{X}$.

following Bellman optimality equations for all $h \in [H]$

$$V^\star_{h,i}(\mathcal{H}_{h,i}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p_h(s' \mid s_{h,i}, a) V^\star_{h+1,i}\left(\mathcal{H}_{h,i} \circ (a, s')\right),$$

with the following boundary conditions: $V^\star_{H+1,i}(\mathcal{H}_{H+1,i}) = V^\star_{1,i+1}(\mathcal{H}_{1,i+1})$ for all $i < K$ and, for the last rollout, $V^\star_{H+1,K}(\mathcal{H}_{H+1,K}) = \max_{j \in [K]} \sum_{h=1}^{H} r_h(s_{h,j}, a_{h,j})$. An optimal deterministic history-dependent policy is obtained by choosing, for every $\mathcal{H}_{h,i}$ with $h \in [H]$, any action that attains the max in the definition of the optimal value function.

3 Planning

We now study the problem of finding an optimal policy for the $\max@k$ problem when $\mathcal{M}$ is known. In Section 3.1, we show that the $\max@k$ criteria change the nature of the problem as Markovian policies are no longer sufficient for optimality. We also prove that the full history is unnecessary for the agent: it suffices knowing the current state, the best return obtained so far, and the cumulative reward in the current rollout. Building on this characterization, in Section 3.2, we propose a polynomial-time approximation scheme for the optimal policy. Finally, in Section 3.3, we show that this is essentially all we can ask for, as we prove that exact planning is computationally hard.

3.1 On the Policy Space

Markovian Policies Are Not Enough. The main challenge in $\max@k$ decision processes is that, unlike the classic MDP setting, we are no longer guaranteed to find optimal policies among Markovian policies, denoted in the following as Π_{M}. Within this context, we refer to Markovian policies as policies that depend on the current state s, the step h, but also the rollout index i.[2] We demonstrate the issue with policies in Π_{M} with a simple example. Consider an MDP, visualized in Figure 1, in which, at a starting state s, there are two actions: one safe, a_s, and one risky, a_r. The safe action leads to a total reward of $\frac{1}{2}$ with probability $1-\gamma$ for some small $\gamma > 0$ (and 0 otherwise), and the risky action leads to a reward of 1 with probability $\frac{1}{4}$ (otherwise it gives a reward of zero).[3] Considering $K = 2$, any Markovian policy will either play a_s in one rollout and a_r in the other, or play both a_s and a_r.[4] It is easy to see that for small γ the best is to play once a_r and once a_s, which gives an expected reward of $\mathbb{E}[\max(\frac{1}{2}\mathsf{Ber}(1-\gamma), \mathsf{Ber}(\frac{1}{4}))] = \frac{5}{8} - \frac{3\gamma}{8}$.[5] On the other hand, the non-Markovian policy which plays a_s in the first rollout and a_r in the second rollout only if the first rollout had a reward of $1/2$ will have an expected reward of $\mathbb{E}[\max(\frac{1}{2}, \mathsf{Ber}(\frac{1}{4}))] = \frac{5}{8}$ conditioned on the first rollout being positive (which happens with probability $1-\gamma$), and $\mathbb{E}[\frac{1}{2}\mathsf{Ber}(1-\gamma)] = \frac{1-\gamma}{2}$ conditioned on the first rollout being negative (which happens with probability γ). In total, this history-dependent policy gives an expected reward of $\frac{5}{8}(1-\gamma) + \gamma\frac{1-\gamma}{2} = \frac{5}{8} - \gamma(\frac{1}{8} + \gamma) > \frac{5}{8} - \frac{3\gamma}{8}$ for small $\gamma > 0$.

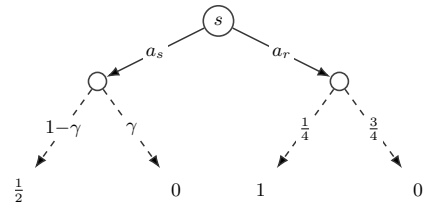

Figure 1: Visualization of the MDP that illustrates the sub-optimality of Markovian policies. Dashed edges denote stochastic transitions, and leaves are the rewards at the end of the horizon.

The Price of Non-Adaptivity We just saw that history-dependent policies Π_{hist} might outperform Markovian ones. We now complement the previous example, with a quantitative analysis of the gap between these two classes of policies. Specifically, we show a multiplicative gap between the performance of history dependent policies and markovian ones is some fixed constant $c > 1$.

[2] One can show stronger sub-optimality by restricting attention to policies that depend only on h and s.

[3] To obtain the reward process described, it is sufficient that, when, *e.g.*, taking a_s, the agent reaches a state where all actions are identical and lead to a leaf with reward $\frac{1}{2}$ and 0 with probability $1-\gamma$ and γ, respectively.

[4] We can ignore stochastic policies as the objective is multilinear in the policy. Hence, there exists a deterministic policies that attains the optimum within Π_{M}.

[5] With $\mathsf{Ber}(p)$ we indicate a Bernoulli random variable with parameter p.

Proposition 1 (Multiplicative Gap of Markovian Policies). *For any $K \geq 2$, there is an instance $\mathcal{M}$ and a state s for which*

$$\max_{\pi \in \Pi_{\text{hist}}} V^{\pi}_{1,1}(s) \geq \left(1.24 - O(\tfrac{1}{K})\right) \cdot \max_{\pi \in \Pi_{\text{M}}} V^{\pi}_{1,1}(s).$$

Recall that a Markovian policy is allowed to depend on the rollout index as well, and thus is a stronger benchmark than using the same Markovian policy repeatedly. This shows that using Markovian policies generally comes at a high cost, and thus it is of paramount importance learning in a larger policy space.

History Can Be Compressed. In the following, we show that it is possible to search for an optimal policy in a space which is smaller than Π_{hist}. Specifically, we prove that at rollout i and step h, it is sufficient to condition $\pi_{h,i}$ on the current state, the maximum return collected in the previous $i-1$ rollouts, and the cumulative reward that has been obtained in the previous $h-1$ steps, at the i-th rollout. To formally state the result, let us first introduce some notation. We denote by $G_{h,i}$ the cumulative reward gathered in rollout i at step h, namely $G_{h,i} = \sum_{k=1}^{h} r_h(s_{h,i}, a_{h,i})$. Furthermore, we denote by $M_i = \max_{j<i} G_{H,j}$ the maximum return collected in the first $i-1$ rollouts, with the convention that $M_0 = 0$. Then, for each step h and rollout i, let $\Phi(\mathcal{H}_{h,i}) = (s_{h,i}, M_i, G_{h,i})$ be a compressed representation of the history and let Π_{C} be the set of policies that, at each step, are conditioned only on the compressed history. Then, we prove the following result.

Proposition 2 (Sufficient conditions for optimality). *In the $\max@k$ problem, it is without loss of optimality to consider only deterministic policies that belong to Π_{C}.*

The proof of Proposition 2 is informative for understanding the problem as it works by construction. Specifically, it shows that there exists a policy $\mu^\star \in \Pi_{\text{C}}$ that attains optimality in terms of Equation (1). To this end, we first prove that value functions of policies in Π_{C} are only a function of the compressed history $\Phi(\mathcal{H}_{h,i})$. As a consequence the value $V^{\mu}_{h,i}(\mathcal{H}_{h,i})$ is constant across histories that have the same shared representation. Furthermore, we also observe that, given two histories with the same Φ's and appending any (a, s') to these histories, we obtain again two histories with the same compressed representation. This, in turn, implies that the optimization problem $\overline{V}^{\star}_{h,i}(\mathcal{H}_{h,i}) \coloneqq \sup_{\mu \in \Pi_{\text{C}}} V^{\mu}_{h,i}(\mathcal{H}_{h,i})$ obeys the following Bellman-like optimality equations

$$\overline{V}^{\star}_{h,i}(\mathcal{H}_{h,i}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p_h(s' \mid s, a) \overline{V}^{\star}_{h+1,i}(\mathcal{H}_{h,i} \circ (a, s')). \tag{2}$$

Then, the last step of the proof is to show by induction that $\mu^\star$ attains the value of the optimal policy among all history-dependent policies, *i.e.*, that $\overline{V}^{\mu^\star} = V^\star$.

Remark. We conclude this section with the following remark. In Proposition 2, we showed that the class of policies Π_{C} is sufficient for optimality. The class of policies Π_{C} depends on the maximum rewards collected at previous rollouts and the current cumulative reward of the rollout. We complement this result by showing that any policy that removes either the current cumulative reward of the maximum return collected at previous rollouts is sub-optimal.

Proposition 3 (Ignoring Previous Best Return or Current Cumulative Return is Sub-optimal). *There exists an instance in which a history-dependent policy gets an expected best return strictly larger than (i) any policy that depends on the current state, rollout index, rollout step, and best return obtained in the previous rollouts, and (ii) any policy that depends on the current state, rollout index, rollout step, and cumulative return in the current rollout.*

The proof of this statement relies on an instance similar to that of Figure 1, except that each rollot begins by randomly receiving a reward of either 0 or 1. Due to space constraints, we defer the high-level intuition behind Proposition 3 to Appendix B. Here, we only mention that this random reward at the beginning of the episode has the purpose of highlighting the role that the current cumulative return plays in defining an optimal behavior.

3.2 Approximate Planning Algorithm

Above, we have shown that, when solving the $\max@k$ problem, one can restrict the search to policies that belong to Π_{C}. However, we observe that solving the optimization problem in this restricted

space, *e.g.*, via backward induction, is computationally demanding. Indeed, suppose to apply backward induction to find the optimal policy $\mu^\star \in \Pi_{\text{C}}$. Then, at the beginning of the induction, we need to consider all the possible compressed history representations $\Phi(\mathcal{H}_{H+1,K})$. The cardinality of this set is $S \cdot |\mathcal{B}| \cdot |\mathcal{C}_{H+1}|$, where $\mathcal{B}$ is the set of possible previous best returns, and $\mathcal{C}_h$ is the set of possible partial returns at step h. More generally, when solving the maximum over actions in Equation (2) at step h in rollout i, we need to consider all the compressed history $\Phi(\mathcal{H}_{h,i})$, whose size is $S \cdot |\mathcal{B}| \cdot |\mathcal{C}_h|$. From this argument, it follows that the complexity of tabular backward induction is of order $\mathcal{O}\left(KAS^2|\mathcal{B}| \sum_h |\mathcal{C}_h|\right)$. The main issue, now, is that $|\mathcal{C}_h|$ and $|\mathcal{B}|$ can be exponential in H.[6] In the following, we discuss how we solve this problem.

The main idea to circumvent this issue is that, intuitively, small changes both in the optimal return observed so far and in the cumulative reward of the current rollout shuold not lead to significant changes in terms of optimal value that the agent can achieve. Therefore, we propose to discretize the compressed history representations $\Phi(\mathcal{H}_{h,i})$ along these two dimensions. To this end, we proceed in the following way. We first discretize the interval $[0, 1]$ of possible rewards using a uniform grid and then project the true reward function onto this set. Precisely, for a given $\kappa > 0$, denote by $\mathcal{R}_\kappa = \{\kappa \cdot (j - \frac{1}{2})\}_{j=1}^{\lceil 1/\kappa \rceil}$ the uniform discretization of the $[0, 1]$ interval. Then, for all state-action pair (s, a) and all steps h, let $\tilde{r}_{h,\kappa}(s, a) = \text{argmin}_{r \in \mathcal{R}_\kappa} |r - r_h(s, a)|$. Once this is done, we can solve by backward induction the $\max@k$ problem for the MDP $\mathcal{M}_\kappa = (\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{\tilde{r}_{h,\kappa}\}_{h\in[H]})$ where we replaced the true reward function with the discretized one. Let us denote by $\tilde{\mu}^\star_\kappa$ the resulting policy for $\mathcal{M}_\kappa$ and observe that $\tilde{\mu}^\star_\kappa$ can be adopted in $\mathcal{M}$ as-is, with the single remark that the compressed history representation uses the discretized rewards rather than the true ones. The following result summarizes the theoretical guarantees of the proposed method.

Theorem 1 (Efficient Approximation Algorithm)**.** *Let $\kappa > 0$ and denote by $\tilde{\mu}^\star_\kappa$ the deterministic optimal policy for $\mathcal{M}_\kappa$ which is found by backward induction on $\mathcal{M}_\kappa$. Then, for any rollout $i \in [K]$, step $h \in [H + 1]$ and history $\mathcal{H}_{h,i}$, it holds that:*

$$|V^\star_{h,i}(\mathcal{H}_{h,i}) - V^{\tilde{\mu}^\star_\kappa}_{h,i}(\mathcal{H}_{h,i})| \leq 2\kappa H. \tag{3}$$

Furthermore, $\tilde{\mu}^\star_\kappa$ can be computed in $\mathcal{O}\left(\frac{KS^2AH^3}{\kappa^2}\right)$ time.

Theorem 1 shows that the proposed method is an *efficient* approximation algorithm for the $\max@k$ objective in the true MDP $\mathcal{M}$. Indeed, setting $\kappa = \epsilon/H$, we have that we can learn an ϵ-optimal policy for $\mathcal{M}$ in $\mathcal{O}(H^5S^2A\epsilon^{-2})$ time, which is polynomial in the size of the instance. The underlying reasoning behind Equation (3) was indeed to show that, for fixed a policy π, evaluated in $\mathcal{M}$ and $\mathcal{M}_\kappa$, yields a performance difference that is at most $H\kappa$ (Lemma 1). Then, a simple error decomposition argument yields Equation (3). Concerning the efficiency of the algorithm, instead, we have that both $|\mathcal{B}|$ and $|\mathcal{C}_h|$ are bounded by $\mathcal{O}(\frac{H}{\kappa})$. Hence, when using backward induction, only a polynomial number of states needs to be considered.

3.3 Impossibility of an Exact Planning Algorithm

In Section 3.2, we showed how to obtain an ϵ-approximation of the planning problem in time inversely polynomial in the approximation error. In this section, we show that this is necessary, and we cannot hope to solve the exact planning problem efficiently. We use tools from computational complexity to establish this lower bound. We reduce from the problem of Subset-Sum, which is the problem of deciding if there exists a subset of natural numbers from a given multiset $S \subset \mathbb{N}$ that sums up to a given threshold $L \in \mathbb{N}$. We consider a decision problem of selecting an optimal action given a certain history.

Theorem 2 (Hardness of Exact Planning)**.** *There is a polynomial time reduction from* Subset-Sum *to the following problem: given $\eta > 0$, an MDP $\mathcal{M}$, a history $\mathcal{H}_{h,i}$ and an action $a \in \mathcal{A}$ decide if $V^\pi_{h,i}(\mathcal{H}_{h,i}) \geq V^{\pi'}_{h,i}(\mathcal{H}_{h,i}) - \eta$ or $V^\pi_{h,i}(\mathcal{H}_{h,i}) < V^{\pi'}_{h,i}(\mathcal{H}_{h,i}) - \eta$, where π is a policy playing deterministically action a conditioned on the history $\mathcal{H}_{h,i}$ and π' is any policy playing deterministically any action different than a.*

[6] To see this, consider an MDP with two actions, and suppose that, at step h, one gives reward 0 and the other 2^{-h}. Then, the cardinality of $|\mathcal{C}_h|$ and $\mathcal{B}$ are clearly 2^h and 2^H respectively.

The practical takeaway from this result is that we cannot hope to find an exact solution to the planning problem efficiently. Therefore, the fully polynomial approximation scheme of Section 3.2 is tight in this sense. Moreover, the gap η used in the proof of Theorem 2 is exponentially small (in the description of the problem). Such a small gap is actually needed as any hardness for a polynomial gap would contradict the existence of the approximation algorithm of Section 3.2.

Interestingly, the proof of Theorem 2 already holds for $K = 2$ and when the set of policies Π_{hist} has constant size. This means that what is truly difficult is not computing an optimal policy but even computing the value of a given one. This shows that that the exact problem is not even in NP. Given the existence of a fully polynomial approximation scheme, our primary interest was to show that no better algorithm could be found, rather than to provide a full characterization of the problem's computational complexity. However, we would like to point out that the Bellman equations reported in Section 2 can likely be turned into a PSPACE membership result. We leave to future work the question of whether this problem is actually complete for PSPACE.

4 Learning

In this section, we study the statistical learning problem associated with the $\max@k$ objective with a generative model. Precisely, we first introduce the PAC framework that we adopt. Then, in Section 4.1, we derive a lower bound for the problem, and finally, in Section 4.2, we propose a uniform sampling algorithm and analyze its theoretical guarantees.

PAC Framework We consider the case where the learner has access to a generative model of the environment and seeks to learn an approximately optimal policy for the $\max@k$ problem in an MDP $\mathcal{M}$ with an unknown transition kernel. The interaction between the agent and the environment is as follows. For a time inhomogeneous MDP, during each learning step $t \in \mathbb{N}$, the agent selects a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and a step $h \in [H]$ and observes a sample $s' \sim p_h(\cdot|s, a)$; for the time homogeneous case, instead, the algorithm simply select (s, a) and observes a sample $s' \sim p(\cdot \mid s, a)$. The learning algorithm takes as input a maximum risk parameter $\delta \in (0, 1)$ and an accuracy level ϵ, and its goal is to output a policy $\hat{\pi} \in \Pi_{\text{C}}$ such that

$$\mathbb{P}\left(V^{\star}_{1,1}(s) - V^{\hat{\pi}}_{1,1}(s) \leq \epsilon,\ \forall s \in \mathcal{S}\right) \geq 1 - \delta, \tag{4}$$

while minimizing the number of interactions with the environment. In words, we want to efficiently find an ϵ-optimal policy with high probability. In the following, we refer to algorithms that satisfy Equation (4) as (ϵ, δ)-correct. Moreover, we denote by τ_δ the stopping time that controls the end of the data acquisition phase, *i.e.*, when the number of samples gathered from the generative model is τ_δ, the algorithm stops and returns the policy $\hat{\pi}$.

Remark on the Generative Model. Studying learning problems in RL with a generative model has a long-standing history in the RL theory community (see, e.g., Gheshlaghi Azar et al., 2013; Sidford et al., 2019; Li et al., 2020; Agarwal et al., 2020; Wang et al., 2021, and references therein). This framework is primarily motivated by the fact that it decouples the estimation problem from the exploration challenges that arise in MDPs. Furthermore, it allows for simpler algorithms and theoretical analyses. In this work, we considered access to a generative model to isolate the estimation problem and understand the level of accuracy required to achieve near-optimal performance. Extending our results to the forward interaction setting (adopted, e.g., in Dann & Brunskill, 2015; Domingues et al., 2021; Zhang et al., 2025b) is an important direction for future work. Nonetheless, as we shall see in this section, adopting a generative model already allows us to understand important features of the $\max@k$ problem and to draw a clear distinction with traditional RL.

4.1 Lower Bound

We start by deriving a lower bound on the statistical complexity of the problem.

Theorem 3 (Statistical Lower Bound)**.** *For every $K \geq 1$, there exists a time-homogeneous MDP instance such that, for any (ϵ, δ)-correct algorithm, it holds that:*

$$\mathbb{E}[\tau_\delta] \geq \Omega\left(\frac{KH^3SA}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right).$$

Results. To better understand Theorem 3, it is helpful to first recall the standard result for the classical RL setting with a generative model. In this latter case, for finite-horizon time-homogeneous MDPs, the lower bound is well-known to be $\Omega\left(H^3SA\epsilon^{-2}\log\left(\frac{1}{\delta}\right)\right)$ (Gheshlaghi Azar et al., 2013; Sidford et al., 2019). This rate is also tight in the minimax sense, meaning that there exists an (ϵ,δ)-correct algorithm that stops with a number of samples that is $\widetilde{\mathcal{O}}\left(H^3SA\epsilon^{-2}\log\left(\frac{1}{\delta}\right)\right)$ (Sidford et al., 2019; Li et al., 2020). Therefore, our Theorem 3 establishes a clear-cut, in terms of statistical efficiency, between standard RL and the $\max@k$ criterion.[7] Specifically, it shows that the number of samples required grows linearly in the number of rollouts K. Furthermore, K actually multiplies the entire rates of standard RL, meaning that the $\max@k$ criterion is, from a statistical viewpoint, at least K times more demanding than classical RL. In words, the takeaway from this result is that although the budget K allows the agent to potentially achieve higher returns by using multiple rollouts, identifying a near-optimal behavior for this objective comes at a non-negligible cost.

The Linear Dependency in K We now provide a sketch behind the argument that allows us to show a linear lower bound in K. The example we are going to discuss is illustrative in that it serves as the building block behind the result of Theorem 3. Indeed, the proof of Theorem 3 essentially follows, from a high-level perspective, the same arguments we present now, with the main difference that the presence of the planning horizon H introduces additional technicalities. Consider the simple instance $\mathcal{M}_0$ depicted in Figure 2. There is a state s with two equal actions a_1 and a_2. Both actions result in a reward of 1 with probability p and 0 with probability $1-p$. Immediately afterward, the rollout terminates and a new one starts from s, so the agent repeat this experiment for K times. Next, consider alternative instances, $\mathcal{M}_1$ and $\mathcal{M}_2$, obtained by making one of the two actions slightly better, *i.e.*, by increasing the success probability from p to $p+\Delta$. Here, both p and Δ should be seen as small numbers, so that the probability of reaching the goal is small in any instance. Then, since the learning algorithm needs to return an ϵ-optimal policy, it should be able to distinguish among these instances. Indeed, suppose that $\mathcal{M}_1$ is true, and consider any policy π. With some algebra and probabilistic arguments, one can show that:

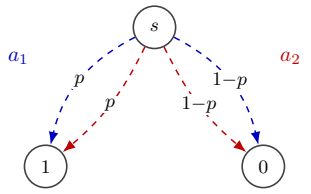

Figure 2: Visualization of the MDP for the linear lower bound in K. There is a state s and two actions, a_1 (blue) and a_2 (red), which both lead to rewards of 1 and 0 with probabilities p and $1-p$.

$$V^{\star}_{1,1}(s) - V^{\pi}_{1,1}(s) \gtrsim \Delta K \sum_{i\in[K]} (1-\alpha_i), \tag{5}$$

where α_i is the probability that π selects a_1 given that the goal was not reached in the i-th rollout.[8] This, in turn, implies that any (ϵ,δ)-correct algorithm should put enough mass on the optimal action a_1 throughout the rollouts, *i.e.*, that $m_1 := \sum_i \alpha_i \gtrsim K - \frac{\Delta}{\epsilon}$. In particular, choosing $\Delta = c\frac{\epsilon}{K}$ for some appropriate constant c, one can show that $m_1 \geq \frac{3}{4}K$, and, thus, for any (ϵ,δ)-correct algorithm it should hold that $\mathbb{P}_{\mathcal{M}_1}(m_1 \geq \frac{3}{4}K) \geq 1-\delta$. Following an analogous argument, we have that, for $\mathcal{M}_2$, $\mathbb{P}_{\mathcal{M}_2}(m_2 \geq \frac{3}{4}K) \geq 1-\delta$, where m_2 is the total mass that the policy puts on a_2 across the different rollouts. Clearly, however, only one of these two events holds, as $m_1 + m_2 = K$. Thus, using change-of-measure arguments, the algorithm must collect enough samples to statistically distinguish whether the Bernoulli parameter of a_1 or a_2 is $p+\Delta$. The information gained from one sample is of order $\mathrm{KL}(p, p+\Delta) \lesssim \Delta^2/p$, and therefore any correct algorithm requires at least order $p\Delta^{-2}\log(1/\delta)$. Choosing $p \approx \frac{1}{K}$ and since $\Delta \approx \frac{\epsilon}{K}$, we obtain a lower bound whose order is $K\epsilon^{-2}\log(1/\delta)$.

Technical Remarks. Before moving to the algorithmic section, we make a remark behind the proof of Theorem 3. The main challenge, compared to the lower bound for PAC-RL, arises directly

[7] As a minor note, we observe that Theorem 3 recovers the standard RL result for the $K=1$ case.

[8] The intuitive reason behind Equation (5) is that what matters in terms of difference of value functions is just the probability with which the policies reach the goal state (*i.e.*, of having one success in K trials). Now, every time that we miss playing a_1, we lose a success probability of Δ. Since p and Δ's are small, the probability that this missed success impacts the value function is bounded away from 0 and, therefore, these losses accumulate over all rollouts in which the policy does not select a_1.

from the $\max@k$ objective and is already present in the illustrative example that we presented above. The point is that, in RL, taking the wrong action immediately results in a loss of return. In $\max@k$, instead, not playing the optimal action has an impact only if the maximum over K rollouts would have changed. Therefore, we need to construct instances in a way that keeps "successes" sufficiently rare so that not playing the optimal action is decisive across several rollouts. This, in turn, also affects the statistical argument. While in standard RL lower bounds, one often argues that the learning agent should identify the unique optimal action in a certain state, here, the agent can distribute its action choices across several rollouts. As a consequence, we needed to develop an argument that accounts for the total mass assigned to the optimal action across several rollouts. The proof must therefore show that ϵ-optimality forces large inter-rollout mass on the good action.

4.2 Algorithm

We now show that there exists an (efficient) algorithm that attains the same rate of Theorem 3. The algorithm combines a standard uniform sampling method (e.g., Gheshlaghi Azar et al., 2013) with the approximation scheme we presented in Section 3. Precisely, the approach works as follows. First, we sample every state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and every step $h \in [H]$ for $m \in \mathbb{N}$ times (for time-homogeneous systems, instead, we only query every state-action pair (s, a) times). The collected samples are used to construct the empirical transition kernel $\{\hat{p}_h\}_{h\in[H]}$. Then, the interval $[0, 1]$ of possible rewards is discretized uniformly using $\mathcal{O}(\kappa^{-1})$ points and a discretized reward function $\{\tilde{r}_{h,\kappa}\}_{h\in[H]}$ is obtained by projecting the true rewards onto the points within the discretization. Once this is done, the algorithm simply stops and return an optimal policy for $\widehat{\mathcal{M}}_\kappa = (\mathcal{S}, \mathcal{A}, s_0, H, \{\hat{p}_h\}_{h\in[H]}, \{\tilde{r}_{h,\kappa}\}_{h\in[H]})$ via backward induction. The following result shows that choosing m and κ appropriately ensures (ϵ, δ)-correctness and minimax optimality.

Theorem 4 (Upper Bound). *Choose* $m = \widetilde{\mathcal{O}}\left(SK^2H^2 \log\left(\frac{1}{\delta}\right) + KH^3 \log(\frac{1}{\delta})\epsilon^{-2}\right)$ *and* $\kappa = \mathcal{O}(\epsilon H^{-1})$. *Then, for any time homogeneous MDP* $\mathcal{M}$, *the proposed algorithm is* (ϵ, δ)*-correct and its sample complexity is given by:*

$$\widetilde{\mathcal{O}}\left(AS^2K^2H^2 \log\left(\frac{1}{\delta}\right) + \frac{KSAH^3}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right). \tag{6}$$

If $\mathcal{M}$ *is time inhomogeneous the sample complexity is* $\widetilde{\mathcal{O}}\left(AS^2K^2H^3 \log\left(\frac{1}{\delta}\right) + \frac{KSAH^4}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right)$.

Results. We start by noticing that, in the small-accuracy regime, the dominant term in Equation (6) is $\widetilde{\mathcal{O}}(KSAH^3/\epsilon^2)$. Therefore, for sufficiently small ϵ, our algorithm matches the lower bound in Theorem 3, thereby establishing tight rates for the $\max@k$ problem. In particular, we have proven that the linear dependency on K is the correct scaling for our setting. As a takeaway, we can conclude that when using a small budget K, the number of samples required to achieve ϵ-optimal behavior is still comparable to that needed for standard RL. The algorithm is also computationally efficient since, after a uniform sampling phase, it solves the empirical discretized problem by dynamic programming on $\widehat{\mathcal{M}}_\kappa$. Due to Theorem 1, this procedure is efficient. Finally, as is usual in PAC learning for MDPs (see, e.g., Li et al., 2020), an additional H-factor is incurred in the sample complexity when the MDP is time-inhomogeneous.

Technical Remarks. To obtain Theorem 4, we first account for the error introduced by planning with discretized rewards. Using arguments similar to Theorem 1 (*i.e.*, small changes in rewards lead to small changes in value functions), this only introduces an additive error term of $\mathcal{O}(H\kappa)$ which we balance by choosing κ as $\mathcal{O}(\epsilon H^{-1})$. The remaining analysis is left to control the planning error that arises from planning with an empirical model. In other words, we need to study how errors propagate within the Bellman equations introduced in Section 3.2. Here, we note that directly applying existing finite-horizon guarantees on the augmented MDP would yield suboptimal dependence on K and H, since the effective horizon is KH. We therefore establish a refined analysis for the simulation error for planning with the empirical model. The techniques behind this are inspired by existing analyses that yield tight rates for classical RL (e.g., Gheshlaghi Azar et al., 2013), but they are tailored to exploit the structure of the $\max@k$ objective to obtain the desired scale in K and H.

AI USE STATEMENT

In this work, generative AI tools were used to assist us with the proof of mathematical claims. In particular, we used AI tools in our research process by querying the generative tools to fill missing details in proof templates that we provided. We also used generative models to check computation and fill in the numerical details of the instances that we used to derive Proposition 3. Additionally, we used generative AI tools for retrieval and discovery of related works in the following sense. After a manual literature review, we integrated in our process generative AI tools to find additional keywords and research areas that were related to our work. Finally, AI tools were used to assist us with the preparation of this document, including writing support and presentation. We have used AI tools to generate the latex code for the figures that are present in this document.

We have reviewed all AI-assisted work. In particular, all the mathematical support provided by generative tools were carefully checked by the authors, and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is primarily theoretical and intended to advance the field of reinforcement learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

REPRODUCIBILITY STATEMENT

This paper is of a theoretical nature. We provide a complete mathematical description of the problem we studied in Section 2. Each theoretical claim is also complemented by a complete proof that can be found in the appendix.

REFERENCES

Mastane Achab, Stephan Clémençon, Aurélien Garivier, Anne Sabourin, and Claire Vernade. Max k-armed bandit: On the extremehunter algorithm and beyond. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 389–404. Springer, 2017.

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

Fahiem Bacchus, Craig Boutilier, and Adam Grove. Rewarding behaviors. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 1160–1167, 1996.

Farid Bagirov, Mikhail Arkhipov, Ksenia Sycheva, Evgeniy Glukhov, and Egor Bogomolov. The best of n worlds: Aligning reinforcement learning with best-of-n sampling via max@ k optimisation. *arXiv preprint arXiv:2510.23393*, 2025.

Amir Rezaei Balef, Claire Vernade, and Katharina Eggensperger. Put cash on bandits: A max k-armed problem for automated machine learning. *arXiv preprint arXiv:2505.05226*, 2025.

Annalisa Barbara, Riccardo Poiani, Martino Bernasconi, and Andrea Celli. Optimal rates for feasible payoff set estimation in games. In *Forty-third International Conference on Machine Learning*, 2026.

Osbert Bastani, Jason Yecheng Ma, Estelle Shen, and Wanqiao Xu. Regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 35:36259–36269, 2022.

Dorian Baudry, Yoan Russac, and Emilie Kaufmann. Efficient algorithms for extreme bandits. *arXiv preprint arXiv:2203.10883*, 2022.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. Pmlr, 2017.

Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning*. MIT Press, 2023.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17682–17690, 2024.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pp. 208–240. Springer, 2003.

Ronen Brafman, Giuseppe De Giacomo, and Fabio Patrizi. Ltlf/ldlf non-markovian rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Alexandra Carpentier and Michal Valko. Extreme bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

Jiacheng Chen, Xinyu Zhang, Shunkai Zhang, Yanmohan Wang, Lin Li, Tiancheng Qin, Qin Wang, Zhengmao Zhu, Tianle Li, Jingyang Li, et al. Maxproof: Scaling mathematical proof with generative-verifier rl and population-level test-time scaling. *arXiv preprint arXiv:2606.13473*, 2026a.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@ k training for adaptively balancing exploration and exploitation of large reasoning models. *arXiv preprint arXiv:2508.10751*, 2025.

Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2026b.

Ziyi Chen, Yan Wen, Zhengmian Hu, and Heng Huang. Robust reinforcement learning with general utility. *Advances in Neural Information Processing Systems*, 37:11290–11344, 2024.

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for cvar optimization in mdps. *Advances in neural information processing systems*, 27, 2014.

Yinlam Chow, Guy Tennenholtz, Izzeddin Gur, Vincent Zhuang, Bo Dai, Aviral Kumar, Rishabh Agarwal, Sridhar Thiagarajan, Craig Boutilier, and Aleksandra Faust. Inference-aware fine-tuning for best-of-n sampling in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 78936–78959, 2025.

Jui-Hui Chung, Ziyang Cai, Zihao Li, Qishuo Yin, Rohit Agarwal, Simon Park, Rodrigo Porto, Narutatsu Ri, Ziran Yang, Shange Tang, et al. Goedel-architect: Streamlining formal theorem proving with blueprint generation and refinement. *arXiv preprint arXiv:2606.06468*, 2026.

Stephen Chung, Wenyu Du, and Jie Fu. Learning from failures in multi-attempt reinforcement learning. *arXiv preprint arXiv:2503.04808*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

Jiaxun Gao, Him Wai Michael Ng, and Z Jane Wang. Once-more: continuous self-correction for large language models via perplexity-guided intervention. In *International Conference on Learning Representations*, volume 2026, pp. 8797–8825, 2026.

Maor Gaon and Ronen Brafman. Reinforcement learning with non-markovian rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3980–3987, 2020.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3): 325–349, 2013.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jinyi Han, Xinyi Wang, Haiquan Zhao, Zishang Jiang, Sihang Jiang, Jiaqing Liang, Xin Lin, Weikang Zhou, Zeye Sun, Fei Yu, et al. A stitch in time saves nine: Proactive self-refinement for language models. In *International Conference on Learning Representations*, volume 2026, pp. 151015–151039, 2026.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International conference on machine learning*, pp. 2681–2691. PMLR, 2019.

Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, 2022.

Muhammed Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, and Samet Oymak. Learning to correct: Calibrated reinforcement learning for multi-attempt chain-of-thought. *arXiv preprint arXiv:2604.17912*, 2026.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Edouard Leurent, and Michal Valko. Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 33:1253–1263, 2020.

Navdeep Kumar, Kaixin Wang, Kfir Levy, and Shie Mannor. Policy gradient for reinforcement learning with general utilities. *arXiv preprint arXiv:2210.00991*, 2022.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in neural information processing systems*, 33:12861–12872, 2020.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Aranyak Mehta, Uri Nadav, Alexandros Psomas, and Aviad Rubinstein. Hitting the high notes: Subset selection for maximizing expected order statistics. *Advances in Neural Information Processing Systems*, 33:15800–15810, 2020.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Challenging common assumptions in convex reinforcement learning. *Advances in Neural Information Processing Systems*, 35:4489–4502, 2022.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Convex reinforcement learning in finite trials. *Journal of Machine Learning Research*, 24(250):1–42, 2023.

Robert Nishihara, David Lopez-Paz, and Léon Bottou. No regret bound for extreme bandits. In *Artificial Intelligence and Statistics*, pp. 259–267. PMLR, 2016.

Soichiro Nishimori, Paavo Parmas, Sotetsu Koyamada, Tadashi Kozuno, Toshinori Kitamura, Shin Ishii, and Yutaka Matsuo. Emergence of exploration in policy gradient reinforcement learning via retrying. *arXiv preprint arXiv:2606.00151*, 2026.

Paavo Parmas, Yongmin Kim, Kohsei Matsutani, Shota Takashiro, Soichiro Nishimori, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Ordergrad: Optimizing beyond the mean with order-statistic policy gradient estimation. *arXiv preprint arXiv:2606.06096*, 2026.

Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

Mark Rowland, Li K Wenliang, Rémi Munos, Clare Lyle, Yunhao Tang, and Will Dabney. Near-minimax-optimal distributional reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 37:132774–132823, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F. Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving discounted markov decision process with a generative model, 2019. URL https://arxiv.org/abs/1806.01492.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 10131–10165, 2025.

Shota Takashiro, Soichiro Nishimori, Paavo Parmas, Yongmin Kim, Kohsei Matsutani, Gouki Minegishi, Yusuke Iwasawa, Takeshi Kojima, and Yutaka Matsuo. On advantage estimates for max@ k policy gradients. *arXiv preprint arXiv:2606.06080*, 2026.

Yunhao Tang, Kunhao Zheng, Gabriel Synnaeve, and Remi Munos. Optimizing language models for inference time objectives using reinforcement learning. In *International Conference on Machine Learning*, pp. 59066–59085. PMLR, 2025.

Bingkui Tong, Junpei Komiyama, Soichiro Nishimori, and Paavo Parmas. Finite-time regret analysis of retry-aware bandits. *arXiv preprint arXiv:2605.20854*, 2026.

Alessandro Trapasso, Luca Iocchi, and Fabio Patrizi. Model-based reinforcement learning in discrete-action non-markovian reward decision processes. *arXiv preprint arXiv:2512.14617*, 2025.

George Tsoukalas, Anton Kovsharov, Sergey Shirobokov, Anja Surina, Moritz Firsching, Gergely Bérczi, Francisco JR Ruiz, Arun Suggala, Adam Zsolt Wagner, Eric Wieser, et al. Advancing mathematics research with ai-driven formal proof search. *arXiv preprint arXiv:2605.22763*, 2026.

Siddarth Venkatraman, Vineet Jain, Sarthak Mittal, Vedant Shah, Johan Obando-Ceron, Yoshua Bengio, Brian R Bartoldson, Bhavya Kailkhura, Guillaume Lajoie, Glen Berseth, et al. Recursive self-aggregation unlocks deep thinking in large language models. *arXiv preprint arXiv:2509.26626*, 2025.

Christian Walder and Deep Tejas Karkhanis. Pass@ k policy optimization: Solving harder reinforcement learning problems. *Advances in Neural Information Processing Systems*, 38:152416–152445, 2026.

Bingyan Wang, Yuling Yan, and Jianqing Fan. Sample-efficient reinforcement learning for linearly-parameterized mdps with a generative model. *Advances in neural information processing systems*, 34:23009–23022, 2021.

Kaiwen Wang, Nathan Kallus, and Wen Sun. Near-minimax-optimal risk-sensitive reinforcement learning with cvar. In *International Conference on Machine Learning*, pp. 35864–35907. PMLR, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zhengqi Wu and Renyuan Xu. Risk-sensitive markov decision process and learning under general utility functions. *arXiv preprint arXiv:2311.13589*, 2023.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. *Advances in Neural Information Processing Systems*, 34:25746–25759, 2021.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.

Liangyu Zhang, Yang Peng, Jiadong Liang, Wenhao Yang, and Zhihua Zhang. Estimation and inference in distributional reinforcement learning. *arXiv preprint arXiv:2309.17262*, 2023.

Shenao Zhang, Yaqing Wang, Yinxiao Liu, Tianqi Liu, Peter Grabowski, Eugene Ie, Zhaoran Wang, and Yunxuan Li. Beyond markovian: Reflective exploration via bayes-adaptive rl for llm reasoning. *arXiv preprint arXiv:2505.20561*, 2025a.

Zihan Zhang, Yuxin Chen, Jason Lee, and Simon S Du. Settling the sample complexity of online reinforcement learning. *Journal of the ACM*, 72(3):1–63, 2025b.

CONTENT

1 Introduction **1**

2 Preliminaries **2**

3 Planning **4**

3.1 On the Policy Space 4

3.2 Approximate Planning Algorithm 5

3.3 Impossibility of an Exact Planning Algorithm 6

4 Learning **7**

4.1 Lower Bound 7

4.2 Algorithm 9

A Additional Related Works **16**

B Proofs of Section 3: Planning **16**

B.1 Proof of Proposition 1 16

B.2 Proof of Proposition 2 17

B.3 Proof of Proposition 3 19

B.4 Proof of Theorem 1 22

B.5 Proof of Theorem 2 23

C Proofs of Section 4: Learning **25**

C.1 Proof of Theorem 3 25

C.2 Proof of Theorem 4 28

C.2.1 Error Decomposition and Simulation Lemmas 29

C.2.2 High Probability Error Bound 32

C.2.3 Putting Everything Together 35

D Auxiliary Lemmas **36**

A Additional Related Works

Convex RL, general utilities, and risk-sensitive objectives From a theoretical perspective, our work is related to RL where the objective is a nonlinear function of the policy-induced occupancy measure (*e.g.*, Hazan et al., 2019; Zhang et al., 2020; Zahavy et al., 2021; Kumar et al., 2022; Mutti et al., 2022; 2023; Chen et al., 2024). These papers are related as nonlinear objectives might invalidate the optimality of Markovian policies. The main difference is that $\max@k$ is not a generic function of the state frequencies but a structured order statistic across several rollouts. Indeed, our study and many of our considerations are tailored around our performance criteria. Another strand of theoretical literature that is close in spirit to our work is risk-sensitive (Chow & Ghavamzadeh, 2014; Fei et al., 2020; Bastani et al., 2022; Wang et al., 2023) and distributional RL (Bellemare et al., 2017; Dabney et al., 2018; Bellemare et al., 2023; Zhang et al., 2023; Rowland et al., 2024; Parmas et al., 2026). Here, however, the focus is primarily either on the problem of precisely learning the distribution of returns or a generic functional of the return distribution. Within this context and close to our work, Wu & Xu (2023) study general utilities of the cumulative reward and recover dynamic programming by augmenting the state with the cumulative reward. Our compressed representation extends this idea to the multi-rollout $\max@k$ setting, where the policy may adapt both across rollouts and within the current rollout. Furthermore, it is worth mentioning the literature that directly studies generic non-Markovian reward decision processes, where the transition dynamics are Markovian but rewards depend on the history (Bacchus et al., 1996; Brafman et al., 2018; Icarte et al., 2022; Gaon & Brafman, 2020; Trapasso et al., 2025). Our setting differs in that the one-step reward is Markovian and history dependence directly arises from the $\max@k$ evaluation criterion.

Retry-aware objectives and extreme bandits Nishimori et al. (2026) and Tong et al. (2026) study retry-aware objectives. Nishimori et al. (2026) focus on RL with the goal of showing that retry-aware optimization can induce stochastic exploration. In this sense, their scope is tailored to answer whether one can avoid explicit exploration bonuses in RL just by changing the objective function to a retry-aware one. In contrast, our work studies the structure of optimal policies under the $\max@k$ criterion and, in particular, how information from previous and current rollouts should be retained for optimal control. Tong et al. (2026), instead, focus on stochastic bandits and provide a finite-time regret analysis of a posterior-sampling rule that selects actions according to the posterior expected maximum over virtual retries. Their setting shares with ours the fact that retry-aware objectives value diversity and high-tail outcomes rather than only mean performance. However, our results differ in two main ways. First, their model is bandit and each attempt is a single arm pull, whereas in our setting each attempt is a finite-horizon trajectory, which introduces sequential decision-making and the structural question of what information from the current and previous rollouts must be retained for optimal control. Second, their guarantees are stated in terms of the standard regret notion for stochastic bandits, while our guarantees are with respect to the $\max@k$ criterion. Another related line of work is extreme bandits, where the classical mean-reward objective is replaced by order-statistic criteria (*i.e.*, objectives that emphasize maxima, tails, or other extreme outcomes) (Carpentier & Valko, 2014; Nishihara et al., 2016). Subsequent work has shown connections between extreme bandits and classical bandit methods (Achab et al., 2017), developed efficient algorithms for this setting using concentration properties of maxima and quantile-of-maxima statistics (Baudry et al., 2022), and has studied applications to model selection (Balef et al., 2025). Finally, we highlight the work of Mehta et al. (2020), which studies subset selection for maximizing expected order statistics and give approximation and hardness results for selecting random variables to maximize the expected maximum or second maximum.

B Proofs of Section 3: Planning

B.1 Proof of Proposition 1

Proof of Proposition 1. Consider any $K \in \mathbb{N}$, and an MDP with a single state and two actions a_s and a_r. Moreover, let $\mu > 1/K$ be a constant defined in the following. a_s leads to a reward of $1/K$ with probability $1/K$ (and 0 otherwise), while a_r leads to a reward of μ with probability $1/K^2$ and 0 otherwise. Consider the following history dependent policy that plays a_s until it observes the first success, then plays a_r. Let W be the number of turns spent by playing a_s by the policy. Conditioning on $W = w \leq K$, the expected reward of this policy is $\frac{1}{K}(1 - \frac{1}{K^2})^{K-w} + \mu(1 - (1 - \frac{1}{K^2})^{K-w})$

and thus

$$\begin{aligned}
\max_{\pi\in\Pi_{\text{hist}}} V^{\pi}_{1,1}(s) &\geq \sum_{w=1}^{K} \tfrac{1}{K}(1-\tfrac{1}{K})^{w-1}\Big[\tfrac{1}{K}(1-\tfrac{1}{K^2})^{K-w} + \mu\big(1-(1-\tfrac{1}{K^2})^{K-w}\big)\Big] \\
&= \tfrac{1}{K^2}\sum_{w=1}^{K}(1-\tfrac{1}{K})^{w-1}(1-\tfrac{1}{K^2})^{K-w} + \tfrac{\mu}{K}\sum_{w=1}^{K}(1-\tfrac{1}{K})^{w-1} - \tfrac{\mu}{K}\sum_{w=1}^{K}(1-\tfrac{1}{K})^{w-1}(1-\tfrac{1}{K^2})^{K-w} \\
&= \tfrac{1}{K}(\tfrac{1}{K}-\mu)\left[\sum_{w=1}^{K}(1-\tfrac{1}{K})^{w-1}(1-\tfrac{1}{K^2})^{K-w}\right] + \tfrac{\mu}{K}\sum_{w=1}^{K}(1-\tfrac{1}{K})^{w-1} \\
&= \tfrac{1}{K}(\tfrac{1}{K}-\mu)\frac{\beta^K-\alpha^K}{\beta-\alpha} + \frac{\mu}{K}\frac{1-\alpha^K}{1-\alpha} \qquad (\alpha = 1-\tfrac{1}{K}, \beta = 1-\tfrac{1}{K^2}) \\
&= (\tfrac{1}{K}-\mu)\frac{K}{K-1}(\beta^K-\alpha^K) + \mu(1-\alpha^K) \\
&= \frac{1}{K-1}(\beta^K-\alpha^K) + \mu\big[1-\tfrac{K}{K-1}\beta^K + \tfrac{1}{K-1}\alpha^K\big].
\end{aligned}$$

Consider now the first term. Clearly $\beta^K \geq 1-1/K$ and $\alpha^K \leq e^{-1}$ and thus the first term can be lower bounded by $(1-1/e)/K - 1/K^2$. For the second term, instead, we can upper bound $\beta^K \leq 1-\frac{1}{K}+\binom{K}{2}\frac{1}{K^4}$ by the binomial expansion and $\frac{1}{K-1}\alpha^K \geq \frac{1}{eK}$ and thus, simplifying, one can lower bound the second term with $\frac{1}{eK} - \frac{1}{2K^2}$. Thus, we obtained that

$$\max_{\pi\in\Pi_{\text{hist}}} V^{\pi}_{1,1}(s) \geq \frac{1-e^{-1}(1-\mu)}{K} - \tfrac{1}{K^2}(1+\tfrac{\mu}{2}).$$

Conversely, consider any deterministic Markovian policy.[9] It is parametrized by the fixed number W of times it will play the safe action a_s (and $K-W$ times it will thus play the risky action a_r). The value is then

$$\begin{aligned}
\max_{\pi\in\Pi_{\text{M}}} V^{\pi}_{1,1}(s) &\leq \max_{W\in[K]}\Big[\mu(1-\beta^{K-W}) + \tfrac{1}{K}\beta^{K-W}(1-\alpha^W)\Big] \\
&\leq \max_{x\in[0,1]}\Big[\tfrac{\mu}{k}(1-x) + \tfrac{1}{k}(1-\alpha^{xK})\Big] \qquad (x := W/k \text{ and } 1-\beta^{xK} \leq (1-x)/K) \\
&\leq \max_{x\in[0,1]}\Big[\tfrac{\mu}{K}(1-x) + \tfrac{1}{k}(1-e^{-x}+K^{-1})\Big],
\end{aligned}$$

which is maximized for $x = \ln(1/\mu)$ and gives

$$\max_{\pi\in\Pi_{\text{M}}} V^{\pi}_{1,1}(s) \leq \tfrac{\mu}{K}(1-\ln(\mu^{-1})) + \tfrac{1}{K}(1-\mu+K^{-1}) = \tfrac{1}{K}(1+\tfrac{1}{K}+\mu\ln(\mu)).$$

Finally, for any $K \geq 2$ and $\mu = 1/2$:

$$\frac{\max_{\pi\in\Pi_{\text{hist}}} V^{\pi}_{1,1}(s)}{\max_{\pi\in\Pi_{\text{M}}} V^{\pi}_{1,1}(s)} \geq \frac{1-e^{-1}(1-\mu)}{1+\mu\ln(\mu)} - O(K^{-1}) \geq 1.24 - O(K^{-1}),$$

concluding the proof. □

B.2 Proof of Proposition 2

Here, we prove that restricting the attention to policies in Π_{C} is sufficient for optimality.

Proof of Proposition 2. We prove the result by construction. Fix any starting $s \in \mathcal{S}$. Specifically, we show that there exists a deterministic policy $\mu^\star \in \Pi_{\text{C}}$ such that for any history $\mathcal{H}_{h,i}$ the following holds:

$$V^{\star}_{h,i}(\mathcal{H}_{h,i}) - V^{\mu^\star}_{h,i}(\mathcal{H}_{h,i}) = 0.$$

[9] Recall that the objective is multi-linear in the policy and hence deterministic policies attains the max in Π_{M}. Therefore, we can focus on deterministic policies and ignore the stochastic ones.

To this end, we first show that value functions of policies in Π_{C} depend on the full history only through the compressed history. Precisely, fix $\mu \in \Pi_{\mathrm{C}}$, and consider two histories $\mathcal{H}_{h,i}^{(1)}$ and $\mathcal{H}_{h,i}^{(2)}$ such that $\Phi(\mathcal{H}_{h,i}^{(1)}) = \Phi(\mathcal{H}_{h,i}^{(2)})$. Then, we prove by induction that $V_{h,i}^{\mu}(\mathcal{H}_{h,i}^{(1)}) = V_{h,i}^{\mu}(\mathcal{H}_{h,i}^{(2)})$. As a base case, consider rollout K and step $H+1$. Since the two histories have the same compressed history, they have the same M_K and the same $G_{H+1,K}$. Hence,

$$V_{H+1,K}^{\mu}(\mathcal{H}_{H+1,K}^{(1)}) = \max\{M_K, G_{H+1,K}\} = V_{H+1,K}^{\mu}(\mathcal{H}_{H+1,K}^{(2)}).$$

Then, suppose that the claim holds for histories at rollout i and step $h+1$, *i.e.*, that for all $\mathcal{H}_{h+1,i}^{(1)}, \mathcal{H}_{h+1,i}^{(2)}$ such that $\Phi(\mathcal{H}_{h+1,i}^{(1)}) = \Phi(\mathcal{H}_{h+1,i}^{(2)})$ we have that $V_{h+1,i}^{\mu}(\mathcal{H}_{h+1,i}^{(1)}) = V_{h+1,i}^{\mu}(\mathcal{H}_{h+1,i}^{(2)})$. We prove that the result holds for histories at rollout i and step $h \in [H]$. First, observe that, since $\Phi(\mathcal{H}_{h,i}^{(1)}) = \Phi(\mathcal{H}_{h,i}^{(2)})$, the two histories have the same current state $s_{h,i}$. Furthermore,

$$\begin{aligned}
V_{h,i}^{\mu}(\mathcal{H}_{h,i}^{(1)}) &= \sum_{a\in\mathcal{A}} \mu_{h,i}(a \mid \mathcal{H}_{h,i}^{(1)}) \sum_{s'\in\mathcal{S}} p_h(s' \mid s_{h,i}, a) V_{h+1,i}^{\mu}(\mathcal{H}_{h,i}^{(1)} \circ (a, s')) \\
&= \sum_{a\in\mathcal{A}} \mu_{h,i}(a \mid \mathcal{H}_{h,i}^{(2)}) \sum_{s'\in\mathcal{S}} p_h(s' \mid s_{h,i}, a) V_{h+1,i}^{\mu}(\mathcal{H}_{h,i}^{(1)} \circ (a, s')) \qquad (\mu \in \Pi_{\mathrm{C}}) \\
&= \sum_{a\in\mathcal{A}} \mu_{h,i}(a \mid \mathcal{H}_{h,i}^{(2)}) \sum_{s'\in\mathcal{S}} p_h(s' \mid s_{h,i}, a) V_{h+1,i}^{\mu}(\mathcal{H}_{h,i}^{(2)} \circ (a, s')) \\
&= V_{h,i}^{\mu}(\mathcal{H}_{h,i}^{(2)}),
\end{aligned}$$

where, in the third equality, we have used the inductive hypothesis. Indeed, observe that $\Phi(\mathcal{H}_{h,i}^{(1)}) = \Phi(\mathcal{H}_{h,i}^{(2)})$ implies that $\Phi(\mathcal{H}_{h,i}^{(1)} \circ (a, s')) = \Phi(\mathcal{H}_{h,i}^{(2)} \circ (a, s'))$ for any state action pair (s', a). It remains to handle the boundary between two consecutive rollouts, but those are direct since $V_{H+1,i}^{\mu}(\mathcal{H}_{H+1,i}) = V_{H+1,i}^{\mu}(\mathcal{H}_{1,i+1})$ by definition. This conclude the inductive argument.

Now, for any history $\mathcal{H}_{h,i}$, let us define the optimal value over policies in Π_{C} as follows:

$$\overline{V}_{h,i}^{\star}(\mathcal{H}_{h,i})(\mathcal{H}_{h,i}) \coloneqq \sup_{\mu\in\Pi_{\mathrm{C}}} V_{h,i}^{\mu}(\mathcal{H}_{h,i}).$$

Since $V_{h,i}^{\mu}(\mathcal{H}_{h,i})$ is constant over all histories with the same compressed representation, we also have that $\overline{V}_{h,i}^{\star}(\mathcal{H}_{h,i})$ is constant among histories with the same compressed representation. Furthermore, since the optimization is only over policies within Π_{C}, we also have that $\overline{V}_{h,i}^{\star}$ satisfies Bellman-like optimality conditions.[10] Specifically, we have that:

$$\overline{V}_{h,i}^{\star}(\mathcal{H}_{h,i})(\mathcal{H}_{h,i}) = \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} p_h(s' \mid s_{h,i}, a) \overline{V}_{h+1,i}^{\star}(\mathcal{H}_{h,i} \circ (a, s')).$$

with boundary conditions $\overline{V}_{H+1,i}^{\star}(\mathcal{H}_{H+1,i}) = \overline{V}_{1,i+1}^{\star}(\mathcal{H}_{1,i+1})$ for all rollouts $i < K$ and, for the last rollout, $\overline{V}_{H+1,K}^{\star}(\mathcal{H}_{H+1,K}) = \max_{j\in[K]} \sum_{h=1}^{H} r_h(s_{h,j}, a_{h,j})$. Then, an optimal deterministic policy $\mu^{\star} \in \Pi_{\mathrm{C}}$ can be found by any action that attains the max in the definition of $\overline{V}^{\star}$.

Finally, in the last step of the proof, we show that $\mu^{\star}$ is optimal in the sense that it attains the same value of the optimal history dependent policy. This is proved with a recursion argument. Specifically, we prove that $\overline{V}_{h,i}^{\star}(\Phi(\mathcal{H}_{h,i})) = V_{h,i}^{\star}(\mathcal{H}_{h,i})$, from which the claim follows directly. As a base case, consider step $H+1$ at rollout K. Observe that

$$\overline{V}_{H+1,K}^{\star}(\Phi(\mathcal{H}_{H+1,K})) = \max_{j\in[K]} \sum_{h=1}^{H} r_h(s_{h,j}, a_{h,j}) = V_{H+1,K}^{\star}(\mathcal{H}_{H+1,K}).$$

[10] To this end, it is important to recall again that $\Phi(\mathcal{H}_{h,i}^{(1)}) = \Phi(\mathcal{H}_{h,i}^{(2)})$ implies that $\Phi(\mathcal{H}_{h,i}^{(1)} \circ (a, s')) = \Phi(\mathcal{H}_{h,i}^{(2)} \circ (a, s'))$ for any state action pair (s', a).

Now, the induction has two parts. First, suppose that the claim holds at step $h+1$ in rollout i, with $h \in [H]$. Then,

$$\begin{aligned}\overline{V}^{\star}_{h,i}(\mathcal{H}_{h,i}) &= \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} p_h(s' \mid s_{h,i}, a)\overline{V}^{\star}_{h+1,i}(\mathcal{H}_{h,i} \circ (s', a)) \\ &= \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} p_h(s' \mid s_{h,i}, a)V^{\star}_{h+1,i}(\mathcal{H}_{h,i} \circ (s', a)) \\ &= V^{\star}_{h,i}(\mathcal{H}_{h,i}).\end{aligned}$$

Second, suppose that $i < K$ and that the claim holds at the beginning of rollout $i+1$. Then, from the boundary conditions on $\overline{V}^{\star}$ and $V^{\star}$, we have that:

$$\overline{V}^{\star}_{H+1,i}(\mathcal{H}_{H+1,i}) = \overline{V}^{\star}_{1,i+1}(\mathcal{H}_{1,i+1}) = V^{\star}_{1,i+1}(\mathcal{H}_{1,i+1}) = V^{\star}_{H+1,i}(\mathcal{H}_{H+1,i}).$$

Combining the base case with the two induction steps proves that $\overline{V}^{\star}_{h,i}(\mathcal{H}_{h,i}) = V^{\star}_{h,i}(\mathcal{H}_{h,i})$. This yields the optimality of $\mu^{\star}$ and, consequently, to the desired result. □

B.3 Proof of Proposition 3

Intuition behind Proposition 3 The proof of Proposition 3 relies on an instance similar to that of Figure 1, except that each rollot begins by randomly receiving a reward of either 0 or 1. The agent then reaches a common decision state $\bar{s}$, where it chooses between a safe action, which adds a deterministic small reward, and a risky action, which adds a larger reward with probability $1/4$ and reward 0 otherwise (see Figure 3). This construction isolates the two pieces of information that are jointly needed for optimality. At $\bar{s}$, the correct local exploration-exploitation tradeoff depends on the current cumulative return G, because the same last action leads to different final rollout returns depending on the progress already made in the current rollout. It also depends on the previous best return M, because the value of taking risk is determined by whether the current rollout only needs to beat a low benchmark or must improve on an already large return. In particular, the optimal policy tolerates high variance when the downside is already protected by M: if the previous best return is large enough, a failed risky attempt does not reduce the $\max@k$ value, while a successful one can still improve it. By contrast, when the current rollout is already on track to produce a competitive return, and the previous best return is low, the optimal policy may avoid unnecessary variance and secure a safe improvement. If the policy observes M but not G, then it knows how good the current rollout must be, but it cannot distinguish the low and high branches within the rollout, and it must take the same last action in situations where the optimal policy takes different actions. Conversely, if the policy observes G but not M, then it can react to the progress of the current rollout, but it must use the same map from G to actions regardless of whether the previous best return is low or high. Therefore, no policy that drops either G or M can implement all optimal continuation decisions simultaneously.

Proof of Proposition 3. The proof is by construction. Consider the $\max@k$ objective with $K = 2$ rollouts and the following finite-horizon MDP (see also Figure 3). There are two actions $\mathcal{A} = \{a_r, a_s\}$, a_r should be interpreted as a risky action, while a_s as a safe one. Each rollout starts in state s_1. At the first step both actions yield 0 reward, and the state transitions to u_0 or u_1, each with probability $\frac{1}{2}$. Then, at the second step, both actions are equivalent: from u_0 the reward is 0, from u_1 the reward is 1, and in both cases the process moves deterministically to a state $\bar{s}$. Once $\bar{s}$ is reached, the safe action a_s provides a deterministic reward of $\frac{1}{2}$, while the risky action a_r gives reward 1 with probability $\frac{1}{4}$ and reward 0 with probability $\frac{3}{4}$. Therefore, at at state $\bar{s}$ the cumulative reward accumulated before choosing the last action is either 0 or 1 each with probability $\frac{1}{2}$. Let G_i be the total return in the i-the i-th rollout, and let G^s_i (resp., G^r_i) be the total return in rollout i if the safe (resp., risky) action is selected. Then, we have

$$\mathbb{P}(G^s_i = \tfrac{1}{2}) = \tfrac{1}{2}, \quad \mathbb{P}(G^s_i = \tfrac{3}{2}) = \tfrac{1}{2}, \quad \mathbb{P}(G^r_i = 0) = \tfrac{3}{8}, \quad \mathbb{P}(G^r_i = 1) = \tfrac{1}{2}, \quad \mathbb{P}(G^r_i = 2) = \tfrac{1}{8}.$$

We denote by $G = \max\{G_1, G_2\}$ the best return among the two rollouts.

Let M be the previous best return (*i.e.*, the return in the first rollout). We define $U^s(M)$ (resp., $U^r(M)$) as the expected $\max@k$ value conditional on entering the second rollout with previous

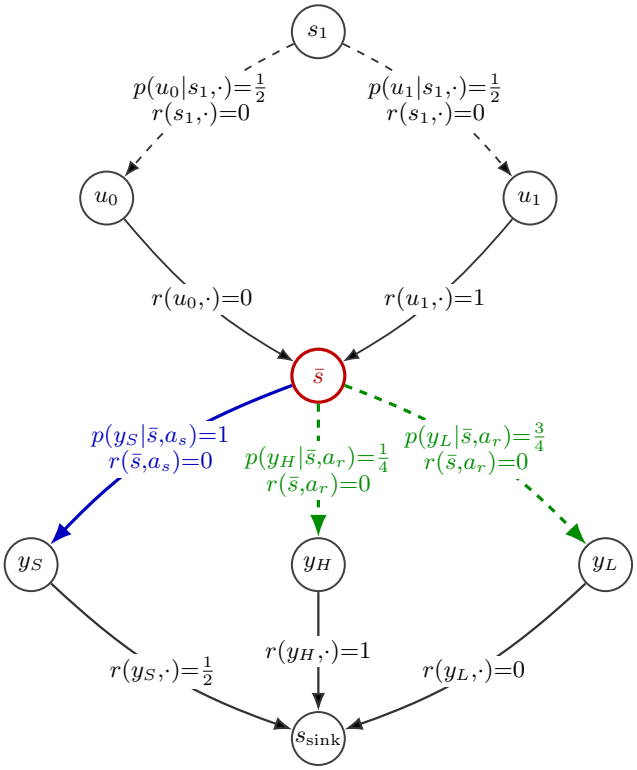

Figure 3: MDP that is used as example to prove Proposition 3.

best return equal to M and playing the safe (resp., risky) action. Using the return distributions computed above we have that

$$U^s(M) := \mathbb{E}[\max\{M, G^s\}] = \frac{1}{2}\max\left\{M, \frac{1}{2}\right\} + \frac{1}{2}\max\left\{M, \frac{3}{2}\right\},$$
$$U^r(M) := \mathbb{E}[\max\{M, G^r\}] = \frac{3}{8}M + \frac{1}{2}\max\{M, 1\} + \frac{1}{8}\max\{M, 2\}.$$

We define the optimal continuation value of a rollout as a function of the previous best return M and the current cumulative return $G \in \{0, 1\}$ at $\bar{s}$. In particular, the two last actions are worth

$$q_s(M, G) = \max\left\{M,\ G + \tfrac{1}{2}\right\}, \qquad q_r(M, G) = \tfrac{1}{4}\max\{M,\ G + 1\} + \tfrac{3}{4}\max\{M,\ G\}.$$

Since $G = 0$ and $G = 1$ each occur with probability $\frac{1}{2}$, the optimal (adaptive) value of a rollout entered with previous best M is

$$U(M) := \frac{1}{2}\max\{q_s(M, 0), q_r(M, 0)\} + \frac{1}{2}\max\{q_s(M, 1), q_r(M, 1)\}.$$

Evaluating U at the only values of M that can arise (a first rollout returns $G_1 \in \{\frac{1}{2}, \frac{3}{2}\}$ under a_s and $G_1 \in \{0, 1, 2\}$ under a_r) gives

$$U(0) = 1, \quad U\left(\frac{1}{2}\right) = \frac{17}{16}, \quad U(1) = \frac{5}{4}, \quad U\left(\frac{3}{2}\right) = \frac{25}{16}, \quad U(2) = 2.$$

We record which action attains the maximum at each of the two branches $G \in \{0, 1\}$ through a pair (a_0, a_1) specifying the action played at $G = 0$ and $G = 1$, respectively. Direct comparison of q_s and q_r yields:

$M = 0$: $q_s > q_r$ at $G = 0$ and $G = 1$, optimal behavior (a_s, a_s);
$M = \frac{1}{2}$: $q_r > q_s$ at $G = 0$, $q_s > q_r$ at $G = 1$, optimal behavior (a_r, a_s);
$M = \frac{3}{2}$: $q_s = q_r$ at $G = 0$, $q_r > q_s$ at $G = 1$, optimal behavior (a_s, a_r) or (a_r, a_r).

These optimal behaviors are pairwise incompatible: no single behavior (a_0, a_1), constant across M, attains $U(M)$ simultaneously at $M = 0$, $M = \frac{1}{2}$, and $M = \frac{3}{2}$.

Lower bound for history-dependent policies. Consider the policy that plays a_s throughout the first rollout, so $G_1 = \frac{1}{2}$ or $G_1 = \frac{3}{2}$ each with probability $\frac{1}{2}$, and then acts greedily in the second rollout. Its value is exactly $\mathbb{E}[U(G_1)]$, so

$$\mathsf{Opt}_{\mathrm{adap}} \;\geq\; \frac{1}{2}\,U\!\left(\frac{1}{2}\right) + \frac{1}{2}\,U\!\left(\frac{3}{2}\right) = \frac{1}{2}\cdot\frac{17}{16} + \frac{1}{2}\cdot\frac{25}{16} = \frac{42}{32}.$$

Policies depending on current state, rollout index, and previous best return. A policy in this class cannot condition on G within the current rollout, so in each rollout it commits to a single last action, independently of G. In the second rollout, entered with previous best M, its value is

$$\max\{U^s(M), U^r(M)\}.$$

In the first rollout the policy commits to a single action. If it plays a_s, then $G_1 \in \{\frac{1}{2}, \frac{3}{2}\}$ each with probability $\frac{1}{2}$, and its value is

$$\frac{1}{2}\max\left\{S\left(\frac{1}{2}\right), R\left(\frac{1}{2}\right)\right\} + \frac{1}{2}\max\left\{S\left(\frac{3}{2}\right), R\left(\frac{3}{2}\right)\right\} = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot\frac{25}{16} = \frac{41}{32}.$$

If it plays a_r, then $G_1 \in \{0, 1, 2\}$ with probabilities $\frac{3}{8}, \frac{1}{2}, \frac{1}{8}$, and its value is

$$\frac{3}{8}\max\{U^s(0), U^r(0)\} + \frac{1}{2}\max\{U^s(1), U^r(1)\} + \frac{1}{8}\max\{U^s(2), U^r(2)\} = \frac{3}{8}\cdot 1 + \frac{1}{2}\cdot\frac{5}{4} + \frac{1}{8}\cdot 2 = \frac{10}{8}.$$

Let $\mathsf{Opt}_{\mathrm{noG}}$ the optimal value attained by the class of policies that cannot condition on G. Taking the best of the two options above we have

$$\mathsf{Opt}_{\mathrm{noG}} = \max\left\{\frac{41}{32}, \frac{10}{8}\right\} = \frac{41}{32} < \mathsf{Opt}_{\mathrm{adap}}.$$

Policies depending on current state, rollout index, and cumulative return. We now consider policies that can condition on the current cumulative return G in the current rollout, but not on the previous best return M. At $\bar{s}$, such a policy selects an action as a function of $G \in \{0, 1\}$ only. Hence, in each rollout its behavior at $\bar{s}$ is described by a pair

$$(a_0, a_1) \in \{a_s, a_r\}^2,$$

where a_g is the action selected when the cumulative return at $\bar{s}$ is $G = g$. We denote the four possible behaviors by

$$ss = (a_s, a_s), \qquad rr = (a_r, a_r), \qquad sr = (a_s, a_r), \qquad rs = (a_r, a_s).$$

Fix any such policy. Let b_1 be its behavior in the first rollout and b_2 its behavior in the second rollout. Since the policy cannot observe the previous best return, b_2 is fixed independently of the realized first-rollout return $M = G_1$. Conditional on M, the value obtained in the second rollout under b_2 is at most $U(M)$, with equality only if the fixed behavior b_2 is optimal for that value of M. Therefore, the value of the policy is bounded by

$$\mathbb{E}\big[U(G_1)\big],$$

with strict inequality whenever b_2 fails to attain $U(M)$ on a value of M that occurs with positive probability under b_1.

We first observe that, if the first-rollout behavior is rr or sr, then

$$\mathbb{E}\big[U(G_1)\big] = \frac{3}{8}U(0) + \frac{1}{2}U(1) + \frac{1}{8}U(2) = \frac{5}{4}$$

for rr, and

$$\mathbb{E}\big[U(G_1)\big] = \frac{1}{2}U\!\left(\frac{1}{2}\right) + \frac{3}{8}U(1) + \frac{1}{8}U(2) = \frac{5}{4}$$

for sr. In both cases, even the adaptive continuation upper bound is strictly smaller than $42/32$. Hence these first-rollout behaviors cannot achieve value $42/32$.

It remains to consider $b_1 = ss$ and $b_1 = rs$. If $b_1 = ss$, then

$$G_1 \in \left\{\frac{1}{2}, \frac{3}{2}\right\}$$

with positive probability on both values. To attain the upper bound $\mathbb{E}[U(G_1)]$, the second-rollout behavior would have to be optimal both at $M = \frac{1}{2}$ and at $M = \frac{3}{2}$. However, from the comparison above, optimality at $M = \frac{1}{2}$ requires behavior (a_r, a_s), whereas optimality at $M = \frac{3}{2}$ requires behavior (a_s, a_r) or (a_r, a_r). No single behavior is optimal for both values of M. Therefore,

$$\mathbb{E}[\max\{G_1, G_2\} \mid b_1 = ss, b_2] < \mathbb{E}\big[U(G_1) \mid b_1 = ss\big] = \frac{1}{2}U\left(\frac{1}{2}\right) + \frac{1}{2}U\left(\frac{3}{2}\right) = \frac{42}{32}.$$

Similarly, if $b_1 = rs$, then

$$G_1 \in \left\{0, 1, \frac{3}{2}\right\},$$

and both 0 and $\frac{3}{2}$ occur with positive probability. To attain the upper bound $\mathbb{E}[U(G_1)]$, the second-rollout behavior would have to be optimal both at $M = 0$ and at $M = \frac{3}{2}$. Optimality at $M = 0$ requires behavior (a_s, a_s), whereas optimality at $M = \frac{3}{2}$ requires behavior (a_s, a_r) or (a_r, a_r). Again, no single behavior is optimal for both values. Hence,

$$\mathbb{E}[\max\{G_1, G_2\} \mid b_1 = rs, b_2] < \mathbb{E}\big[U(G_1) \mid b_1 = rs\big] = \frac{3}{8}U(0) + \frac{1}{8}U(1) + \frac{1}{2}U\left(\frac{3}{2}\right) = \frac{42}{32}.$$

Combining the four cases for b_1, every policy that conditions on the current state, rollout index, and current cumulative return, but not on the previous best return, has value $\mathsf{Opt}_{\mathrm{noM}}$ strictly smaller than $42/32$. Since we already exhibited a history-dependent policy with value at least $42/32$, we obtain

$$\mathsf{Opt}_{\mathrm{noM}} < \mathsf{Opt}_{\mathrm{adap}}.$$

This proves the result. □

B.4 Proof of Theorem 1

To prove Theorem 1 we first derive the following result that studies the difference, in term of performance, of a given history dependent policy π which is evaluated on two different MDPs that differs only in term of the reward function.

Lemma 1 (Value Difference: Same Policy, Different Rewards)**.** *Consider two MDPs $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ that differs only in the reward function and denote by $\{r_h)\}_{h=1}^H$ and $\{\widetilde{r}_h)\}_{h=1}^H$ the reward functions of $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ respectively. Consider any history dependent policy $\pi \in \Pi_{\text{hist}}$. Then, for any rollout $i \in [K]$, step h, and history $\mathcal{H}_{h,i}$ it holds that:*

$$\left|V^{\pi}_{h,i}(\mathcal{H}_{h,i}) - \widetilde{V}^{\pi}_{h,i}(\mathcal{H}_{h,i})\right| \le H\|r - \widetilde{r}\|_{\infty},$$

where $V^{\mu}_{h,i}(\cdot)$ and $\widetilde{V}^{\mu}_{h,i}(\cdot)$ denotes the value functions of policy π in the MDPs $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ respectively.

Proof. We prove this result by induction. As a base case, consider step $H+1$ at rollout K. Here, for any history $\mathcal{H}_{H+1,K}$ it holds that:

$$\begin{aligned}
\left|V^{\pi}_{H+1,K}(\mathcal{H}_{H+1,K}) - \widetilde{V}^{\pi}_{H+1,K}(\mathcal{H}_{H+1,K})\right| &= \left|\max_{j\in[K]} \sum_{h\in[H]} r_h(s_{h,j}, a_{h,j}) - \max_{j\in[K]} \sum_{h\in[H]} \widetilde{r}_h(s_{h,j}, a_{h,j})\right| \\
&\le \sum_{h\in[H]} \max_{s,a,h} |r_h(s,a) - \widetilde{r}_h(s,a)| \\
&= H\|r - \widetilde{r}\|_{\infty}
\end{aligned}$$

Next, suppose that the claims holds at rollout i and step $h+1$. Then, we have that:

$$\begin{aligned}\left|V^{\pi}_{h,i}(\mathcal{H}_{h,i}) - \widetilde{V}^{\pi}_{h,i}(\mathcal{H}_{h,i})\right| &= \Big|\sum_{a\in\mathcal{A}} \pi_{h,i}(a \mid \mathcal{H}_{h,i}) \sum_{s'\in\mathcal{S}} p_h(s' \mid s,a)\left(V^{\pi}_{h+1,i}(\mathcal{H}_{h,i}\circ(a,s')) - \widetilde{V}^{\pi}_{h+1,i}(\mathcal{H}_{h,i}\circ(a,s'))\right)\Big| \\ &\leq H\|r-\widetilde{r}\|_\infty \Big|\sum_{a\in\mathcal{A}} \pi_{h,i}(a \mid \mathcal{H}_{h,i}) \sum_{s'\in\mathcal{S}} p_h(s' \mid s,a)\Big| \quad \text{(Inductive hp.)} \\ &= H\|r-\widetilde{r}\|_\infty.\end{aligned}$$

Finally, suppose that $i < K$ and that the claims holds at the beginning of rollout $i+1$. Then, from the boundary conditions of the value function, we have that:

$$\left|V^{\pi}_{h,i}(\mathcal{H}_{H+1,i}) - \widetilde{V}^{\pi}_{H+1,i}(\mathcal{H}_{H+1,i})\right| = \left|V^{\pi}_{1,i+1}(\mathcal{H}_{1,i+1}) - \widetilde{V}^{\pi}_{1,i+1}(\mathcal{H}_{1,i+1})\right| \leq H\|r-\widetilde{r}\|_\infty.$$

This concludes the proof. □

We now prove Theorem 1.

Proof of Theorem 1. We first prove Equation (3) by induction. To formally prove it, we need to introduce some auxiliary notation. Fix $\kappa > 0$; then, for any history dependent policy $\pi \in \Pi_{\text{hist}}$, we write $V^{\pi,\kappa}$ for the value function of policy π in $\mathcal{M}_\kappa$. Now, let $\mu^\star$ be the optimal policy in Π_{C} for $\mathcal{M}$, and let $\tilde{\mu}^\star_\kappa \in \Pi^\kappa_{\text{C}}$ be the optimal policy for $\mathcal{M}_\kappa$. It holds that:

$$\begin{aligned}V^{\mu^\star}_{h,i}(\mathcal{H}_{h,i}) - V^{\tilde{\mu}^\star_\kappa}_{h,i}(\mathcal{H}_{h,i}) &= V^{\mu^\star}_{h,i}(\mathcal{H}_{h,i}) - V^{\tilde{\mu}^\star_\kappa,\kappa}_{h,i}(\mathcal{H}_{h,i}) + V^{\tilde{\mu}^\star_\kappa,\kappa}_{h,i}(\mathcal{H}_{h,i}) - V^{\tilde{\mu}^\star_\kappa}_{h,i}(\mathcal{H}_{h,i}) \\ &\leq V^{\mu^\star}_{h,i}(\mathcal{H}_{h,i}) - V^{\mu^\star,\kappa}_{h,i}(\mathcal{H}_{h,i}) + V^{\tilde{\mu}^\star_\kappa,\kappa}_{h,i}(\mathcal{H}_{h,i}) - V^{\tilde{\mu}^\star_\kappa}_{h,i}(\mathcal{H}_{h,i}) \\ &\leq 2H\|r-\tilde{r}\|_\infty \qquad \text{(Lemma 1)} \\ &\leq 2H\kappa, \qquad \text{(Def. of } \kappa)\end{aligned}$$

where the first inequality follows from the fact that $\tilde{\mu}^\star_\kappa$ is an optimal policy for for $\mathcal{M}_\kappa$. This concludes the proof of Equation (3).

Finally, we conclude by analyzing the computational complexity of solving $\mathcal{M}_\kappa$. At the first step of the backward induction, the compressed and discretized history representations are $|\mathcal{B}_\kappa|$ and $|\mathcal{C}_{H+1,\kappa}|$, where $\mathcal{B}_\kappa$ is the set of possible previous best returns in the discretized MDP $\mathcal{M}_\kappa$, and $\mathcal{C}_{h,\kappa}$ is the set of possible partial returns at step $h \in [H+1]$ in $\mathcal{M}_\kappa$. More generally, when solving the maximum over actions in the backward induction at step h and rollout i, the set of discretized compressed histories has size $S \cdot |\mathcal{B}_\kappa| \cdot |\mathcal{C}_{h,\kappa}|$. From this argument, it follows that the complexity of tabular backward induction is of order $\mathcal{O}\left(KAS^2|\mathcal{B}_\kappa|\sum_{h=1}^{H+1}|\mathcal{C}_{h,\kappa}|\right)$. Indeed, for each of the K rollouts, at step h there are $|C_{h,\kappa}| \cdot |\mathcal{B}_\kappa| \cdot S$ compressed history representations, and the per-step complexity of this method is SA. Now, observe that both $|\mathcal{B}_\kappa|$ and $|C_{h,\kappa}|$ can be bounded by $1 + \frac{H}{\kappa}$. This yields the result. □

B.5 Proof of Theorem 2

Proof of Theorem 2. We are given a Subset-Sum instance $S = w_1, \ldots, w_n$, and target value L. We define a few parameters first. Let $\eta = \frac{2^{-n}}{8}$, a normalization constant $C = L + 5 + \sum_{j=1}^n w_j$, and a threshold $\beta = \frac{2^{-n}}{16C}$.

Construction. We build the following MDP. We consider only $K = 2$ rollouts. We start from the state s_0, in which we have three actions a^{safe}, a^+, and a^-.

- Playing a^{safe} leads deterministically to a reward of $(1+L)/C$ and the state transition deterministically to an absorbing state.
- Playing a^+ leads to an immediate return of $(1+\eta)/C$ after which the state transition to the state s_1.

- Action a^- leads to an immediate reward of 0 with probability $1/2$ and of $2/C$ with probability $1/2$, and in both cases it transitions to the state s_1;
- At each state s_i with $i \in \{1, \dots, n\}$, there is only a dummy action a_0, which collects with probability $1/2$ a reward of w_i/C and 0 otherwise, and transits deterministically to the state s_{i+1}. State s_n, with probability $1/2$ collects a reward of $(L+3)/C$ and 0 otherwise, after which the game ends.

Consider the random variable: $X = \sum_{j=1}^{n} w_j Z_j + (L+3)Z_0$, where $Z_0, Z_1, \dots, Z_n$ are i.i.d. Bernoulli of parameter $1/2$. There are essentially only three actions in the MDP. Playing a^{safe} leads to the deterministic reward of $R^{\text{safe}} = (1+L)/C$. Playing a^+ leads to a stochastic reward of $R^+ = (1+\eta+X)/C$, while playing a^- leads to a stochastic reward of $R^- = (X+2Z)/C$, where Z is a Bernoulli of parameter $1/2$.

Useful Claims. The main property that we need in this reduction is that $\mathbb{P}(X = L) > 0$ if and only if the SUBSET-SUM instance is in the language, i.e., if there exists $S^* \subseteq S$ such that $\sum_{j \in S^*} w_j = L$. In particular, if there exists such a set S^* we get that

$$\mathbb{P}[X = L] \geq \left(\frac{1}{2}\right)^{n+1}. \tag{7}$$

Then define Q^+ (Q^- respectively), the expected values of playing a^{safe} in the first rollout and a^+ (a^- respectively) in the second rollout. Formally:

$$Q^+ = \mathbb{E}[\max(R^+ - R^{\text{safe}}, 0)] \quad \text{and} \quad Q^- = \mathbb{E}[\max(R^- - R^{\text{safe}}, 0)].$$

Moreover, notice that

$$\begin{aligned} Q^+ &= \frac{1}{C}\mathbb{E}[\max(\eta + X - L, 0)] \\ &= \frac{1}{C}\mathbb{E}[\max(X - L, 0)] + \frac{1}{C}\eta\mathbb{P}[X \geq L]. \end{aligned}$$

where the last equality follows from $\eta < 1$ and X, L are natural numbers. Similarly, one can verify that:

$$\begin{aligned} Q^+ &= \frac{1}{C}\mathbb{E}[\max(X - L + 2Z - 1, 0)] \\ &= \frac{1}{C}\mathbb{E}[\max(X - L, 0)] + \frac{1}{2C}\mathbb{P}[X = L]. \end{aligned}$$

Notably, we can conclude that $Q^- - Q^+ = \frac{1}{C}\left(\frac{1}{2}\mathbb{P}[X = L] - \eta\mathbb{P}[X \geq T]\right)$.

Completeness. Assume that there exists a set $S^* \subseteq S$ such that $\sum_{j \in S^*} w_j = L$. Assume we played a^{safe} in the first rollout. Now we are given the history of the first rollout, and we have to determine the best action to play. Playing either a^+ or a^- is optimal because, with some probability, we will get a large payoff of $(T+3)/C$ in state s_n, which will dominate the reward of $R^{\text{safe}} = (L+1)/C$, collected in the first rollout. Since we are in a Yes instance, we know that:

$$\begin{aligned} Q^- - Q^+ &= \frac{1}{C}\left(\frac{1}{2}\mathbb{P}[X = L] - \eta\mathbb{P}[X \geq L]\right) \\ &\geq \frac{1}{C}\left(\frac{1}{2}\left(\frac{1}{2}\right)^{n+1} - \eta\right) && \text{(Equation (7) and } \mathbb{P}[X \geq L] \leq 1) \\ &= \frac{1}{C}\frac{2^{-n}}{8} \geq \beta, \end{aligned}$$

showing that playing a^- is better with a gap of β than playing a^+ in this history.

Soundness. Consider the case in which there is no such subset S^*. Then $\mathbb{P}(X = L) = 0$. However, we can always get more than $(L+1)/C$ since at state s_n we can receive a reward of $(L+3)/C$ with probability $1/2$, i.e., $\mathbb{P}(X > L) \geq 1/2$. Hence

$$Q^- - Q^+ = -\frac{\eta}{C}\mathbb{P}[X \geq L] \leq -\frac{\eta}{C}\frac{1}{2} = -\frac{2^{-n}}{16C} = -\beta.$$

showing that playing a^- is worst with a gap of β than playing a^+ in this history. □

C PROOFS OF SECTION 4: LEARNING

C.1 PROOF OF THEOREM 3

Proof of Theorem 3. The proof of Theorem 3 is articulated in four steps. First, we present the family of instances that we consider to derive the results. In words, we introduce a base instance, in which all actions are identical, and then, for each state-action pair, we introduce an alternative instance where that action is slightly better for that state. In the second step, we show that in any alternative instance any ϵ-optimal policy should put sufficient mass on the optimal action for that instance. Finally, in the third step, we use this result to define a good event, which we use in the last step (in combination with change-of-measure arguments) to lower bound the expected stopping time of any (ϵ, δ)-correct algorithm.

Step 1: Instances Construction The base instance $\mathcal{M}_0 = (\mathcal{S}, \mathcal{A}, p^{(0)}, r, H)$ is as follows. The set of states is $\mathcal{S} = \{g\} \cup \overline{\mathcal{S}}$, where g is a rewarding and absorbing goal state, and $\overline{\mathcal{S}} = \{s_1, \dots, s_n\}$ is a set of states from which we can try to reach the goal. For the goal state, we have $p^{(0)}(g \mid g, a) = 1$, and $r(g, a) = 1$ for all actions $\forall a \in \mathcal{A}$. For any other state $s_i \in \overline{\mathcal{S}}$ and action $a \in \mathcal{A}$, instead, we have $r(s_i, a) = 0$ and $p^{(0)}(g \mid s_i, a) = p$ and $p^{(0)}(s_i \mid s_i, a) = 1 - p$, where $p = c_0 \frac{1}{KH}$ for some positive constant $c_0 \in \mathbb{R}_{>0}$ which is defined below. Next, for each state-action pair $(s, a) \in \overline{\mathcal{S}} \times \mathcal{A}$, we introduce an alternative instance $\mathcal{M}_{s,a} = (\mathcal{S}, \mathcal{A}, p^{(s,a)}, r, H)$ which is equal to $\mathcal{M}_0$ everywhere but in (s, a). In particular, $p^{(s,a)}(g \mid s, a) = p + \Delta$ and $p^{(s,a)}(s \mid s, a) = 1 - p - \Delta$, where $\Delta = c_1 \frac{\epsilon}{KH^2}$ and $c_1 \in \mathbb{R}_{>0}$ is a positive constant that is defined below. In the rest of the proof, we denote by $\mathfrak{M} = \mathcal{M}_0 \cup \{\mathcal{M}_{s,a}\}_{s \in \overline{\mathcal{S}}, a \in \mathcal{A}}$ the family of instances that we are considering.

We choose $c_0 = 1/96$ and $c_1 = 8192$. Under the assumption that $\epsilon \leq \frac{H}{96 \cdot 8196}$, we have that $\Delta \leq \frac{p}{2}$. Furthermore, we also observe that $p \leq \frac{1}{4}$ for any choice of K, H so that all transitions are correctly defined. In the rest of the proof, we also make the assumption that H is divisible by 4 to simplify some computations. Finally, we suppose that $A \geq 2$ and $\delta \leq \frac{1}{16}$.

Step 2: Value Function Analysis Fix any $\mathcal{M}_{s,a} \in \mathfrak{M}$ and let $\pi \in \Pi_{\mathrm{C}}$ and let $L = \frac{H}{4}$. Since $\mathcal{M}_{s,a}$ is fixed within step, we drop the dependency on $\mathcal{M}_{s,a}$ in symbols such as $\mathbb{E}$ and $\mathbb{P}$. With some algebraic manipulations, we have that:

$$
\begin{aligned}
V^\star_{1,1}(s) - V^\pi_{1,1}(s) &= V^\star_{1,1}(s) - \mathbb{E}_\pi \left[V^\pi_{H,K}(\mathcal{H}_{H,K}) \right] \\
&= V^\star_{1,1}(s) - \mathbb{E}_\pi \left[\max_{i \in [K]} \sum_{h \in [H]} r(s_{h,i}, a_{h,i}) \mid \mathcal{H}_{H,K} \right] \\
&= \sum_{i \in [K]} \sum_{h \in [H]} \mathbb{E}_\pi \left[V^*_{h,i}(\mathcal{H}_{h,i}) - V^\star_{h^+,i^+}(\mathcal{H}_{h^+,i^+}) \right] \qquad \text{(Telescoping argument)} \\
&\geq \sum_{i \in [K]} \sum_{h \in [L]} \mathbb{E}_\pi \left[V^\star_{h,i}(\mathcal{H}_{h,i}) - V^\star_{h^+,i^+}(\mathcal{H}_{h^+,i^+}) \right],
\end{aligned}
$$

where the last step follows from the fact that $\mathbb{E}_\pi[V^\star_{h,i}(\mathcal{H}_{h,i})] \geq \mathbb{E}_\pi[V^\star_{h^+,i^+}(\mathcal{H}_{h^+,i^+})]$. Indeed,

$$
\begin{aligned}
\mathbb{E}_\pi \left[V^\star_{h,i}(\mathcal{H}_{h,i}) \right] &= \mathbb{E}_\pi \left[\max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V^\star_{h^+,i^*}(\mathcal{H}_{h,i} \circ (s', a)) \right] \\
&\geq \mathbb{E}_\pi \left[\sum_{a \in \mathcal{A}} \pi_{h,i}(a \mid \mathcal{H}_{h,i}) \sum_{s' \in \mathcal{S}} p(s' \mid s_{h,i}, a) V^\star_{h^+,i^*}(\mathcal{H}_{h,i} \circ (s', a)) \right] \\
&= \mathbb{E}_\pi [V^\star_{h^+,i^+}(\mathcal{H}_{h^+,i^+})].
\end{aligned}
$$

Now, for any pair (h, i), let us denote by $\mathcal{F}_{h,i}$ the event that, for all $(\bar{h}, \bar{i}) \prec (h, i)$, $s_{h,i} \neq g$. Then, we have that:

$$
V^\star_{1,1}(s) - V^\pi_{1,1}(s) \geq \sum_{i \in [K]} \sum_{h \in [L]} \mathbb{E}_\pi \left[\left(V^\star_{h,i}(\mathcal{H}_{h,i}) - V^\star_{h^+,i^+}(\mathcal{H}_{h^+,i^+}) \right) \mathbb{1}\{\mathcal{F}_{h,i}\} \right]. \tag{8}
$$

We now analyze a single term within the r.h.s of Equation (8). To this end, we first make some consideration. First of all, we note that (a) on $\mathcal{F}_{h,i}$ the state in $\mathcal{H}_{h,i}$ is $s_{h,i} = s$ and (b) the optimal policy always play action a in state s. Furthermore, (c) we observe that $V^\star(\mathcal{H}_{h,i})$ does not depend on the identity of the actions that have been played throughout the history, but only on the current state, the reward cumulated in the current rollouts, and the best return observed so far. Therefore, in the following, we simplify the notation and write $V^\star_{h^+,i^+}(\mathcal{H}_{h,i} \circ \bar{s})$ instead of $V^\star_{h^+,i^+}(\mathcal{H}_{h,i} \circ (\bar{s}, \bar{a}))$ for any state-action pair $(\bar{s}, \bar{a}) \in \mathcal{S} \times \mathcal{A}$. Finally, since we are considering $\pi \in \Pi_{\mathrm{C}}$ and since we are interested only in the case where $\mathcal{F}_{h,i}$ holds, we observe that $\pi_{h,i}(\cdot \mid \mathcal{H}_{h,i})$ is uniquely determined by the pair (h, i).[11] Therefore, we will shrink the notation and write $\alpha^{\pi,s,a}_{h,i} := \pi_{h,i}(a \mid \mathcal{H}_{h,i} \wedge \mathcal{F}_{h,i})$. Here, the dependency on s is due to the fact, on $\mathcal{F}_{h,i}$ the state is necessarily s (see point (a) above. To ease the notation, since (s, a) is fixed in this step, we will drop the dependency on s, a within $\alpha^{\pi,s,a}_{h,i}$. At this point, we can proceed by analyzing the r.h.s. of Equation (8). We first focus on $\mathbb{E}_\pi \left[V^\star_{h,i}(\mathcal{H}_{h,i}) \mathbb{1}\{\mathcal{F}_{h,i}\} \right]$. We have that:

$$\mathbb{E}_\pi \left[V^\star_{h,i}(\mathcal{H}_{h,i}) \mathbb{1}\{\mathcal{F}_{h,i}\} \right] = \mathbb{E}_\pi \left[\Big((p+\Delta) V^\star_{h^+,i^+}(\mathcal{H}_{h,i} \circ g) + (1-p-\Delta) V^\star_{h^+,i^+}(\mathcal{H}_{h,i} \circ s) \Big) \mathbb{1}\{\mathcal{F}_{h,i}\} \right],$$

where the result is a direct application of the definition of value function. For the second term of the r.h.s in Equation (8), namely $\mathbb{E}_\pi \left[V^\star_{h^+,i^+}(\mathcal{H}_{h^+,i^+}) \mathbb{1}\{\mathcal{F}_{h,i}\} \right]$, instead, we have that:

$$\begin{aligned}
\mathbb{E}_\pi \left[V^\star_{h^+,i^+}(\mathcal{H}_{h^+,i^+}) \mathbb{1}\{\mathcal{F}_{h,i}\} \right] &= \mathbb{E}_\pi \left[\mathbb{E}_\pi \left[V^\star_{h^+,i^+}(\mathcal{H}_{h^+,i^+}) \mathbb{1}\{\mathcal{F}_{h,i}\} \mid \mathcal{H}_{h,i} \right] \right] && \text{(Tower law)} \\
&= \mathbb{E}_\pi \left[\Big((p + \Delta\alpha^\pi_{h,i}) V^\star_{h^+,i^+}(\mathcal{H}_{h,i} \circ g) + (1-p-\Delta\alpha^\pi_{h,i}) V^\star_{h^+,i^+}(\mathcal{H}_{h,i} \circ s) \Big) \mathbb{1}\{\mathcal{F}_{h,i}\} \right].
\end{aligned}$$

where, in the second step, we have exploited the definition of $\mathcal{M}_{s,a}$ together with the fact that the value function of $V^\star$ does not depend on the identity of the previously played actions but only on the state, the value of the current rollout, and the best return collected so far. Plugging in these results within Equation (8), we have that:

$$V^\star_{1,1}(s) - V^\pi_{1,1}(s) \geq \sum_{i \in [K]} \sum_{h \in [L]} \mathbb{E}_\pi \left[\Delta(1 - \alpha^\pi_{h,i}) \Big(V^\star_{h^+,i^+}(\mathcal{H}_{h,i} \circ g) - V^\star_{h^+,i^+}(\mathcal{H}_{h,i} \circ s) \Big) \mathbb{1}\{\mathcal{F}_{h,i}\} \right]. \tag{9}$$

Now, we lower bound and upper bound $V^\star_{h^+,i^+}(\mathcal{H}_{h,i} \circ g)$ and $V^\star_{h^+,i^+}(\mathcal{H}_{h,i} \circ s)$. For the first of these two terms, we can use that the agent reached the goal sufficiently early since $h \in [L]$. Thus, we have that:

$$V^\star_{h^+,i^+}(\mathcal{H}_{h,i} \circ g) \geq H - L \geq \frac{3}{4} H.$$

For the second term, instead, we exploit the fact that the goal is an hard-to-reach state:

$$V^\star_{h^+,i^+}(\mathcal{H}_{h,i} \circ s) \leq V^\star_{1,1}(s) \leq H \mathbb{P}_{\pi^\star}(\exists (h,i) : s_{h,i} = g) \leq K H^2 (p + \Delta) \leq \frac{H}{64}$$

[11] Indeed, as we noted above, the state is necessarily s, the previous best return is 0, and the cumulative reward in the current rollout is 0 as well.

where the last step follows from the definition of p and the fact that $c_0 \leq 1/96$. Plugging these results within Equation (9) and using some probabilistic arguments we obtain:

$$
\begin{aligned}
V^{\star}_{1,1}(s) - V^{\pi}_{1,1}(s) &\geq \frac{47\Delta H}{64} \sum_{i\in[K]} \sum_{h\in[L]} (1-\alpha^{\pi}_{h,i}) \mathbb{P}_{\pi}\left(\forall(\bar{h},\bar{i}) \prec (h,i) : s_{\bar{h},\bar{i}} \neq g\right) \\
&= \frac{47\Delta H}{64} \sum_{i\in[K]} \sum_{h\in[L]} (1-\alpha^{\pi}_{h,i}) \prod_{(\bar{h},\bar{i})\prec(i,h)} \mathbb{P}_{\pi}\left(s_{\bar{h},\bar{i}} \neq g \mid \mathcal{F}_{\bar{h}^-,\bar{i}^-}\right) \quad \text{(Chain rule)} \\
&= \frac{47\Delta H}{64} \sum_{i\in[K]} \sum_{h\in[L]} (1-\alpha^{\pi}_{h,i}) \prod_{(\bar{h},\bar{i})\prec(i,h)} \left(\alpha^{\pi}_{\bar{i}\bar{h}}\left(1-p-\Delta\right) + (1-\alpha^{\pi}_{\bar{h},\bar{i}}(1-p)\right)) \\
&\geq \frac{47\Delta H}{64} \sum_{i\in[K]} \sum_{h\in[L]} (1-\alpha^{\pi}_{h,i}) (1-p-\Delta)^{KH} \\
&\geq \frac{1}{4} \cdot \frac{47\Delta H}{64} \left(KL - \sum_{i\in[K]} \sum_{h\in[L]} \alpha^{\pi}_{h,i} \right) \\
&\qquad \text{(Bernoulli's inequality, def. of } p, \Delta \text{ and } c_0 \leq 1/96)
\end{aligned}
$$

Therefore, rearranging the terms, we have that, for any ϵ-optimal policy in $\mathcal{M}_{s,a}$ it must hold that:

$$
\begin{aligned}
\sum_{i\in[K]} \sum_{h\in[L]} \alpha^{\pi}_{h,i} &\geq KL - \frac{256\epsilon}{47\Delta H} \\
&= KH\left(\frac{1}{4} - \frac{256}{47c_1}\right) && \text{(Def. of } \Delta) \\
&\geq \frac{KL}{2}. && (c_1 \geq 8192)
\end{aligned}
$$

Step 3: Good Event In the previous step, we have shown that, for any $\mathcal{M}_{s,a}$, an ϵ-optimal policy must satisfy $\sum_{i\in[K]} \sum_{h\in[L]} \alpha^{\pi,s,a}_{h,i} \geq \frac{KL}{2}$. Therefore, for the policy $\hat{\pi}$ that is returned by an (ϵ,δ)-correct algorithm on $\mathcal{M}_{s,a}$ must satisfy this requirement in high-probability. Precisely, for each $(s,a) \in \overline{S} \times A$, we denote by define $\mathcal{E}_{s,a}$ as follows:

$$
\mathcal{E}_{s,a} \coloneqq \left\{ \sum_{i\in[K]} \sum_{h\in[L]} \alpha^{\hat{\pi},s,a}_{h,i} \geq \tfrac{KL}{2} \right\}.
$$

Since the algorithm is (ϵ,δ)-correct, it must hold that

$$
\mathbb{P}_{\mathcal{M}_{s,a}}(\mathcal{E}_{s,a}) \geq 1-\delta. \tag{10}
$$

Step 4: Change of Measure Finally, we combine the previous results with change-of-measure arguments. Let us denote by $N_{s,a}(\tau_\delta) = \sum_{t=1}^{\tau_\delta} \mathbb{1}\{s_t = s, a_t = a\}$ the number of samples gathered by algorithm at state-action pair (s,a) up to time τ_δ. Then, for $\mathcal{M}_{s,a} \in \mathcal{M}$. For any random any random-variable Z that takes value in $[0,1]$, we have that:

$$
\begin{aligned}
\mathrm{kl}(\mathbb{E}_{\mathcal{M}_0}[Z], \mathbb{E}_{\mathcal{M}_{s,a}}[Z]) &\leq \mathrm{KL}\left(\mathbb{P}_{\mathcal{M}_0}, \mathbb{P}_{\mathcal{M}_{s,a}}\right) && \text{(Lemma 11)} \\
&\leq \mathbb{E}_{\mathcal{M}_0}[N_{s,a}(\tau_\delta)] \cdot \mathrm{KL}(p^{(0)}(\cdot \mid s,a), p^{(s,a)}(\cdot \mid s,a)) \\
&\leq \mathbb{E}_{\mathcal{M}_0}[N_{s,a}(\tau_\delta)] \cdot \frac{\mathrm{TV}(p^{(0)}(\cdot \mid s,a), p^{(s,a)}(\cdot \mid s,a))^2}{(p+\Delta)} \\
&\qquad \text{(Reverse Pinsker's Inequality)} \\
&= \mathbb{E}_{\mathcal{M}_0}[N_{s,a}(\tau_\delta)] \cdot \frac{8\Delta^2}{p},
\end{aligned}
$$

where $\mathrm{kl}(p,q)$ denotes the KL-divergence between Bernoulli distributions with mean p and q and $\mathrm{KL}\left(\mathbb{P}_{\mathcal{M}_0}, \mathbb{P}_{\mathcal{M}_{s,a}}\right)$ denotes the KL divergence between $\mathbb{P}_{\mathcal{M}_0}$ and $\mathbb{P}_{\mathcal{M}_{s,a}}$. Next, we apply the previ-

ous result using $Z = \mathbb{1}\{\mathcal{E}_{s,a}\}$ and we obtain that:

$$\begin{aligned}
\mathbb{E}_{\mathcal{M}_0}[\tau_\delta] &\geq \sum_{s\in\overline{S}}\sum_{a\in\mathcal{A}} \mathbb{E}_{\mathcal{M}_0}[N_{s,a}(\tau_\delta)] \\
&\geq \frac{p}{8\Delta^2}\sum_{s\in\overline{S}}\sum_{a\in\mathcal{A}} \mathrm{kl}\left(\mathbb{P}_{\mathcal{M}_0}(\mathcal{E}_{s,a}), \mathbb{P}_{\mathcal{M}_{s,a}}(\mathcal{E}_{s,a})\right) \\
&\geq \frac{p}{8\Delta^2}\sum_{s\in\overline{S}}\sum_{a\in\mathcal{A}} \left(1 - \mathbb{P}_{\mathcal{M}_0}\left(\mathcal{E}_{s,a}\right)\right)\log\left(\frac{1}{1-\mathbb{P}_{\mathcal{M}_{s,a}}(\mathcal{E}_{s,a})}\right) - \log(2) \qquad \text{(Lemma 12)} \\
&\geq \frac{p}{2\Delta^2}\left(\overline{S}A\log\left(\frac{1}{\delta}\right) - \log(2)\overline{S}A - \log\left(\frac{1}{\delta}\right)\sum_{s\in\overline{S}}\sum_{a\in\mathcal{A}}\mathbb{P}_{\mathcal{M}_0}\left(\mathcal{E}_{s,a}\right)\right). \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Equation (10))}
\end{aligned}$$

At this point, let us analyze in more detail $\sum_{s\in\overline{S}}\sum_{a\in\mathcal{A}}\mathbb{P}_{\mathcal{M}_0}(\mathcal{E}_{s,a})$. Observe, that, for any policy π and any state s, it holds that

$$\sum_{a\in\mathcal{A}}\sum_{i\in[K]}\sum_{h\in[L]}\alpha_{h,i}^{\pi,s,a} = KL. \tag{11}$$

Therefore, it holds that:

$$\begin{aligned}
\sum_{s\in\overline{S}}\sum_{a\in\mathcal{A}}\mathbb{P}_{\mathcal{M}_0}(\mathcal{E}_{s,a}) &= \sum_{s\in\overline{S}}\sum_{a\in\mathcal{A}}\mathbb{E}_{\mathcal{M}_0}\left[\mathbb{1}\left\{\sum_{i\in[K]}\sum_{h\in[L]}\alpha_{h,i}^{\hat{\pi},s,a} \geq \tfrac{KL}{2}\right\}\right] \\
&= \sum_{s\in\overline{S}}\mathbb{E}_{\mathcal{M}_0}\left[\sum_{a\in\mathcal{A}}\mathbb{1}\left\{\sum_{i\in[K]}\sum_{h\in[L]}\alpha_{h,i}^{\hat{\pi},s,a} \geq \tfrac{KL}{2}\right\}\right] \\
&\leq \overline{S},
\end{aligned}$$

where in the inequality step we have used Equation (11). Thus, we obtained:

$$\begin{aligned}
\mathbb{E}_{\mathcal{M}_0}[\tau_\delta] &\geq \frac{p}{2\Delta^2}\left(\overline{S}A\log\left(\frac{1}{\delta}\right) - \log(2)\overline{S}A - \log\left(\frac{1}{\delta}\right)\overline{S}\right) \\
&\geq \frac{p}{16\Delta^2}SA\log\left(\frac{1}{\delta}\right) \qquad (A, S \geq 2 \text{ and } \delta \leq \tfrac{1}{16}) \\
&= \frac{c_0}{8c_1^2}\cdot\frac{KH^3AS}{\epsilon^2}\log\left(\frac{1}{\delta}\right),
\end{aligned}$$

which concludes the proof. □

C.2 Proof of Theorem 4

Proof roadmap. The proof proceeds in four main steps. First, Lemma 2 separates the learning error into a reward-discretization term, of order $H\kappa$, and two value-estimation terms for the discretized MDPs $\mathcal{M}_\kappa$ and $\widehat{\mathcal{M}}_\kappa$. Second, we reduce these two MDPs to finite-horizon MDPs on the augmented state space $\mathcal{X}$, whose state records the current original state, the best return obtained in previous rollouts, and the cumulative reward in the current rollout. This reduction allows us to compare $\mathcal{M}_\kappa$ and $\widehat{\mathcal{M}}_\kappa$ only through their transition kernels. Third, Lemma 3 and Lemma 4 give simulation bounds for the value-estimation errors: the first controls the discretized optimal policy $\mu_\kappa^\star$, while the second controls the data-dependent empirical optimizer $\hat{\mu}$ through the reference value function of $\mu_\kappa^\star$. Combining these two bounds yields the refined decomposition in Lemma 5, which reduces the problem to bounding the quantities $C_{\mu_\kappa^\star,\mu_\kappa^\star}$ and $C_{\hat{\mu},\mu_\kappa^\star}$. These quantities measure the accumulated one-step transition-estimation error along the trajectory distribution induced by the first policy in the subscript, where the empirical transition error is tested against the continuation value function of the second policy. So $C_{\mu_\kappa^\star,\mu_\kappa^\star}$ controls the estimation error for the fixed discretized optimizer, whereas $C_{\hat{\mu},\mu_\kappa^\star}$ controls the data-dependent empirical optimizer through the reference value function $V^{\mu_\kappa^\star,\kappa}$.

Fourth, Lemma 6 proves that the cumulative variance of the reference value function along any policy is at most $3H^2$, and Lemma 7 combines this variance estimate with concentration for the empirical transition kernel to obtain a high-probability bound of order $\sqrt{KH^3/m} + KH^2/m + H\kappa$. The theorem then follows by choosing $\kappa = \epsilon/(9H)$ and m large enough so that the statistical and discretization errors are each absorbed into the target accuracy ϵ.

C.2.1 Error Decomposition and Simulation Lemmas

We start by introducing some additional notation used throughout the section. For any fixed $\kappa > 0$ and $m \in \mathbb{N}$, we denote by $\hat{\mu}$ the optimal policy for $\widehat{\mathcal{M}}_\kappa = (\mathcal{S}, \mathcal{A}, s_0, H, \{\hat{p}_h\}_{h\in[H]}, \{\tilde{r}_{h,\kappa}\}_{h\in[H]})$. We recall that $\hat{\mu}$ is the policy returned by the algorithm we presented in Section 4.2. Furthermore, we denote by $\pi^\star$ the optimal policy for $\mathcal{M}$, and by $\mu^\star_\kappa$ the optimal policy for $\mathcal{M}_\kappa$. For any policy π, let V^π and $\widehat{V}^\pi$ denote the value functions of π in $\mathcal{M}$ and $\widehat{\mathcal{M}}$, respectively. Similarly, let $V^{\pi,\kappa}$ and $\widehat{V}^{\pi,\kappa}$ denote the value functions of π in $\mathcal{M}_\kappa$ and $\widehat{\mathcal{M}}_\kappa$, respectively.

The following error decomposition shows that it suffices to control value-estimation errors for the discretized-reward MDPs.

Lemma 2 (Error Decomposition)**.** *It holds that:*

$$V^\star - V^{\hat{\mu}} \le V^{\mu^\star_\kappa,\kappa} - V^{\hat{\mu},\kappa} + 3H\kappa \le \left(V^{\mu^\star_\kappa,\kappa} - \widehat{V}^{\mu^\star_\kappa,\kappa}\right) + \left(\widehat{V}^{\hat{\mu},\kappa} - V^{\hat{\mu},\kappa}\right) + 3H\kappa.$$

Proof. With algebraic manipulations, we have that:

$$\begin{aligned}
V^\star - V^{\hat{\mu}} &= \left(V^\star - V^{\mu^\star_\kappa,\kappa}\right) + \left(V^{\mu^\star_\kappa,\kappa} - V^{\hat{\mu},\kappa}\right) + \left(V^{\hat{\mu},\kappa} - V^{\hat{\mu}}\right) \\
&\le V^{\mu^\star_\kappa,\kappa} - V^{\hat{\mu},\kappa} + 3H\kappa && \text{(Theorem 1 + Lemma 1)} \\
&= V^{\mu^\star_\kappa,\kappa} - \widehat{V}^{\hat{\mu},\kappa} + \widehat{V}^{\hat{\mu},\kappa} - V^{\hat{\mu},\kappa} + 3H\kappa \\
&\le \left(V^{\mu^\star_\kappa,\kappa} - \widehat{V}^{\mu^\star_\kappa,\kappa}\right) + \left(\widehat{V}^{\hat{\mu},\kappa} - V^{\hat{\mu},\kappa}\right) + 3H\kappa, && (\hat{\mu} \text{ is optimal for } \widehat{\mathcal{M}}_\kappa)
\end{aligned}$$

which concludes the proof. □

Reduction to finite horizon MDPs Given Lemma 2, the two quantities of interest are $V^{\mu^\star_\kappa,\kappa} - \widehat{V}^{\mu^\star_\kappa,\kappa}$ and $\widehat{V}^{\hat{\mu},\kappa} - V^{\hat{\mu},\kappa}$. Both terms involve only value functions of policies that belong to Π_{C} evaluated in MDPs that use the discretized reward function, *i.e.*, $\mathcal{M}_\kappa$ and $\widehat{\mathcal{M}}_\kappa$. Observe that these MDPs differ only in the transition kernel. As a consequence, in order to bound these terms, it is convenient to reduce $\mathcal{M}_\kappa$ and $\widehat{\mathcal{M}}_\kappa$ to (almost) standard finite-horizon MDPs. Let $\mathcal{X} := \mathcal{S} \times \lceil \frac{H}{\kappa} \rceil \times \lceil \frac{H}{\kappa} \rceil$ be an extended set of states such that each state $x = (s, M, G) \in \mathcal{X}$ can be interpreted as follows: $s \in \mathcal{S}$ is a state of the original state space, M denotes the best return observed in the previous rollouts, and G denotes the current cumulative reward in the current rollout. The action set $\mathcal{A}$ remains the same. Then, for each triple $(x, a, x') \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$, we define the new transition kernel as follows:

$$\begin{aligned}
T_{h,i}(x' \mid x, a) &= p_h(s' \mid s, a)\mathbb{1}\{M' = M\}\mathbb{1}\{G' = G + \tilde{r}_{h,\kappa}(s, a)\} \quad \forall h < H, \forall i \in [K] \\
T_{H,i}(x' \mid x, a) &= \mathbb{1}\{s' = s_1\}\mathbb{1}\{M' = \max\{M, G + \tilde{r}_{H,\kappa}(s, a)\}\}\mathbb{1}\{G' = 0\} \quad \forall i \in [K].
\end{aligned}$$

As one can see, the state evolves according to (1) the original transition kernel p and (2) the discretized reward function $\tilde{r}$ so that M and G preserve the meaning of the $\max@k$ objective. Finally, we define the new reward function as follows:

$$\begin{aligned}
R_{h,i}(x, a) &= 0 \quad \forall (h, i) \neq (H, K) \\
R_{H,K}(x, a) &= \max\{M, G + \tilde{r}_H(s, a)\}.
\end{aligned}$$

At this point, we turned the original problem into a traditional finite-horizon one where the goal is to find a policy that maximizes the standard RL objective in the MDP

$$\mathcal{M}^{\text{ext}}_\kappa := \left(\mathcal{X}, \mathcal{A}, (s_1, 0, 0), KH, \{T_{h,i}\}_{h\in[H],i\in[K]}, \{R_{h,i}\}_{h\in[H],i\in[K]}\right),$$

where $(s_1, 0, 0) \in \mathcal{X}$ is the initial state in the extended state space.

Given this reduction, in the following, we treat the value functions of policies in Π_{C} as $|\mathcal{X}|$-dimensional vectors (once the step h and the rollout i are fixed). Next, to simplify the notation, it will be convenient to write (h^+, i^+) to denote the (step,rollout) pair that follows (h, i). Precisely, we have that:

$$(h^+, i^+) = \begin{cases} (h+1, i), & \text{if } h < H, \\ (1, i+1), & \text{if } h = H. \end{cases}$$

Moreover, for any two pair (h_1, i_1) and (h_2, i_2), we write $(h_1, i_1) \preceq (h_2, i_2)$ if (h_1, i_1) precedes (h_2, i_2). Furthermore, given any policy $\mu \in \Pi_{\mathrm{C}}$, step h and rollout i, we introduce the policy-induced transition kernel $T^{\mu}_{h,i}$, that is defined as $T^{\mu}_{h,i}(x' \mid x) = \sum_{a \in \mathcal{A}} \mu_{h,i}(a \mid x)\, T_{h,i}(x' \mid x, a)$. In this way, we can compactly express the value function as follows:

$$V^{\mu,\kappa}_{h,i} = T^{\mu}_{h,i} V^{\mu,\kappa}_{h^+,i^+}, \qquad \forall (h,i) \neq (H,K),$$
$$V^{\mu,\kappa}_{H,K}(x) = V^{\mu,\kappa}_{H,K} = \sum_{a \in \mathcal{A}} \mu_{H,K}(a \mid x) \max\{M, G + \tilde{r}_{H,\kappa}(s,a)\} \qquad \forall x = (s, M, G).$$

Finally, before continuing, we extend this construction to the case in which we use an approximation $\hat{p}$ of p. In this case, we denote the extended MDP with $\widehat{\mathcal{M}}^{\mathrm{ext}}_{\kappa}$. This object differs from $\mathcal{M}^{\mathrm{ext}}_{\kappa}$ only in the transition kernel, where T is replaced with its empirical estimate $\widehat{T}$.

Now, we can proceed with the proof. First we introduce some notation. For any pair of policies $\mu, \nu \in \Pi_{\mathrm{C}}$, we define, for brevity, $D_{\mu}, D_{\mu,\nu}$ as follows:

$$D_{\mu} \coloneqq \max_{\bar{h},\bar{i},\bar{x}} \|\widehat{T}^{\mu}_{\bar{h},\bar{i}}(\cdot \mid \bar{x}) - T^{\mu}_{\bar{h},\bar{i}}(\cdot \mid \bar{x})\|_1 \max_{\bar{h},\bar{i}} \|V^{\mu,\kappa}_{\bar{h},\bar{i}} - \widehat{V}^{\mu,\kappa}_{\bar{h},\bar{i}}\|_{\infty}$$
$$D_{\mu,\nu} \coloneqq \max_{\bar{h},\bar{i},\bar{x}} \|\widehat{T}^{\mu}_{\bar{h},\bar{i}}(\cdot \mid \bar{x}) - T^{\mu}_{\bar{h},\bar{i}}(\cdot \mid \bar{x})\|_1 \max_{\bar{h},\bar{i}} \|V^{\mu,\kappa}_{\bar{h},\bar{i}} - V^{\nu,\kappa}_{\bar{h},\bar{i}}\|_{\infty}.$$

Moreover, for the sake of brevity, for a policy μ and a fixed (h, i, x), we denote $\mathbb{E}_{\mu}[\cdot \mid X_{h,i} = x]$ by $\mathbb{E}^{x}_{\mu}[\cdot]$. At this point, we can proceed by analyzing the error terms that appeared in Lemma 2. We first prove two "simulation lemmas" that will be used to control the error terms of Lemma 2.

Lemma 3 (Simulation Lemma 1)**.** *Consider any policy $\mu \in \Pi_{\mathrm{C}}$. Suppose that:*

$$\max_{h,i,x} \|\widehat{T}^{\mu}_{h,i}(\cdot \mid x) - T^{\mu}_{h,i}(\cdot \mid x)\|_1 \leq \frac{1}{2KH}. \tag{12}$$

Then, it holds that:

$$\max_{h,i} \|V^{\mu,\kappa}_{h,i} - \widehat{V}^{\mu,\kappa}_{h,i}\|_{\infty} \leq 2 \max_{h,i,x} \sum_{(\bar{h},\bar{i}) \succeq (h,i)} \mathbb{E}^{x}_{\mu}\left[\left| \sum_{x' \in \mathcal{X}} (T^{\mu}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) - \widehat{T}^{\mu}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}})) V^{\mu,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right| \right].$$

Proof. The lemma easily follows from the definition of value functions at (H, K). We now prove it for $(h, i) \neq (H, K)$. Specifically:

$$\begin{aligned} \left| V^{\mu,\kappa}_{h,i} - \widehat{V}^{\mu,\kappa}_{h,i} \right| &= \left| T^{\mu}_{h,i} V^{\mu,\kappa}_{h^+,i^+} - \widehat{T}^{\mu}_{h,i} \widehat{V}^{\mu,\kappa}_{h^+,i^+} \right| \\ &\leq \left| (T^{\mu}_{h,i} - \widehat{T}^{\mu}_{h,i}) V^{\mu,\kappa}_{h^+,i^+} \right| + T^{\mu}_{h,i} \left| V^{\mu,\kappa}_{h^+,i^+} - \widehat{V}^{\mu,\kappa}_{h^+,i^+} \right| + \left| (\widehat{T}^{\mu}_{h,i} - T^{\mu}_{h,i})(V^{\mu,\kappa}_{h^+,i^+} - \widehat{V}^{\mu,\kappa}_{h^+,i^+}) \right| \\ &\leq \left| (T^{\mu}_{h,i} - \widehat{T}^{\mu}_{h,i}) V^{\mu,\kappa}_{h^+,i^+} \right| + T^{\mu}_{h,i} \left| V^{\mu,\kappa}_{h^+,i^+} - \widehat{V}^{\mu,\kappa}_{h^+,i^+} \right| + D_{\mu} \end{aligned}$$

Now, fix a state $x \in \mathcal{X}$ and analyze the upper bound on $\left| V^{\mu,\kappa}_{h,i}(x) - \widehat{V}^{\mu,\kappa}_{h,i}(x) \right|$ that we just obtained. Using the previous equation we have that:

$$\left| V^{\mu,\kappa}_{h,i}(x) - \widehat{V}^{\mu,\kappa}_{h,i}(x) \right| = \left| \sum_{x' \in \mathcal{X}} (T^{\mu}_{h,i}(x' \mid x) - \widehat{T}^{\mu}_{h,i}(x' \mid x)) V^{\mu,\kappa}_{h^+,i^+}(x') \right| + \mathbb{E}_{T^{\mu}_{h,i}(x)} \left[\left| V^{\mu,\kappa}_{h^+,i^+}(x') - \widehat{V}^{\mu,\kappa}_{h^+,i^+}(x') \right| \right] + D_{\mu}.$$

Unrolling the the r.h.s. of this equation, we have that:

$$\left| V^{\mu,\kappa}_{h,i}(x) - \widehat{V}^{\mu,\kappa}_{h,i}(x) \right| \leq \sum_{(\bar{h},\bar{i}) \succeq (h,i)} \mathbb{E}^{x}_{\mu}\left[\left| \sum_{x' \in \mathcal{X}} (T^{\mu}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) - \widehat{T}^{\mu}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}})) V^{\mu,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right| \right] + KHD_{\mu}.$$

Now, taking maximum on both sides, using the definition of D_{μ} together with Equation (12), and re-arranging the terms yield the result. □

Next, we continue with the second simulation lemma, which bounds the value estimation error for a policy μ using the value function of a reference policy ν.

Lemma 4 (Simulation Lemma 2). *Consider any policy $\mu, \nu \in \Pi_{\mathrm{C}}$. Suppose that:*

$$\max_{h,i,x} \|\widehat{T}^{\mu}_{h,i}(\cdot \mid x) - T^{\mu}_{h,i}(\cdot \mid x)\|_1 \leq \frac{1}{2KH}. \tag{13}$$

It holds that:

$$\max_{h,i} \|V^{\mu,\kappa}_{h,i} - \widehat{V}^{\mu,\kappa}_{h,i}\|_\infty \leq 2 \max_{h,i,x} \sum_{(\bar{h},\bar{i}) \succeq (h,i)} \mathbb{E}^{x}_{\mu}\left[\left| \sum_{x' \in \mathcal{X}} (T^{\mu}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) - \widehat{T}^{\mu}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}})) V^{\nu,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right|\right] + 2KHD_{\mu,\nu}.$$

Proof. The lemma easily follows from the definition of value functions at (H, K). Now, we prove it for $(h, i) \neq (H, K)$. We have that:

$$\begin{aligned}
\left|V^{\mu,\kappa}_{h,i} - \widehat{V}^{\mu,\kappa}_{h,i}\right| &= \left|T^{\mu}_{h,i} V^{\mu,\kappa}_{h^+,i^+} - \widehat{T}^{\mu}_{h,i} \widehat{V}^{\mu,\kappa}_{h^+,i^+}\right| \\
&\leq \left|(T^{\mu}_{h,i} - \widehat{T}^{\mu}_{h,i}) V^{\mu,\kappa}_{h^+,i^+}\right| + T^{\mu}_{h,i}\left|V^{\mu,\kappa}_{h^+,i^+} - \widehat{V}^{\mu,\kappa}_{h^+,i^+}\right| + \left|(\widehat{T}^{\mu}_{h,i} - T^{\mu}_{h,i})(V^{\mu,\kappa}_{h^+,i^+} - \widehat{V}^{\mu,\kappa}_{h^+,i^+})\right| \\
&\leq \left|(T^{\mu}_{h,i} - \widehat{T}^{\mu}_{h,i}) V^{\mu,\kappa}_{h^+,i^+}\right| + T^{\mu}_{h,i}\left|V^{\mu,\kappa}_{h^+,i^+} - \widehat{V}^{\mu,\kappa}_{h^+,i^+}\right| + D_{\mu} \\
&\leq \left|(T^{\mu}_{h,i} - \widehat{T}^{\mu}_{h,i}) V^{\nu,\kappa}_{h^+,i^+}\right| + \left|(T^{\mu}_{h,i} - \widehat{T}^{\mu}_{h,i})(V^{\mu,\kappa}_{h^+,i^+} - V^{\nu,\kappa}_{h^+,i^+})\right| + T^{\mu}_{h,i}\left|V^{\mu,\kappa}_{h^+,i^+} - \widehat{V}^{\mu,\kappa}_{h^+,i^+}\right| + D_{\mu} \\
&\leq \left|(T^{\mu}_{h,i} - \widehat{T}^{\mu}_{h,i}) V^{\nu,\kappa}_{h^+,i^+}\right| + T^{\mu}_{h,i}\left|V^{\mu,\kappa}_{h^+,i^+} - \widehat{V}^{\mu,\kappa}_{h^+,i^+}\right| + D_{\mu} + D_{\mu,\nu}.
\end{aligned}$$

Now, fix a state $x \in \mathcal{X}$ and analyze the upper bound on $\left|V^{\mu,\kappa}_{h,i}(x) - \widehat{V}^{\mu,\kappa}_{h,i}(x)\right|$ that we just obtained. Let $D \coloneqq D_{\mu} + D_{\mu,\nu}$. Then, we can rewrite the previous equation as:

$$\left|V^{\mu,\kappa}_{h,i}(x) - \widehat{V}^{\mu,\kappa}_{h,i}(x)\right| = \left| \sum_{x' \in \mathcal{X}} (T^{\mu}_{h,i}(x' \mid x) - \widehat{T}^{\mu}_{h,i}(x' \mid x)) V^{\nu,\kappa}_{h^+,i^+}(x') \right| + \mathbb{E}_{T^{\mu}_{h,i}(x)}\left[\left|V^{\mu,\kappa}_{h^+,i^+}(x') - \widehat{V}^{\mu,\kappa}_{h^+,i^+}(x')\right|\right] + D.$$

Unrolling the the r.h.s. of this equation, we have that:

$$\left|V^{\mu,\kappa}_{h,i}(x) - \widehat{V}^{\mu,\kappa}_{h,i}(x)\right| \leq \sum_{(\bar{h},\bar{i}) \succeq (h,i)} \mathbb{E}^{x}_{\mu}\left[\left| \sum_{x' \in \mathcal{X}} (T^{\mu}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) - \widehat{T}^{\mu}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}})) V^{\nu,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right|\right] + KHD.$$

Now, taking maximum on both sides, using the definition of D_{μ} together with Equation (13) and re-arranging the terms yield the result. □

We now use Lemma 3 and Lemma 4 to obtained a refined version of Lemma 2.

Lemma 5 (Refined Error Decomposition). *For any pair of policies $\mu, \nu \in \Pi_{\mathrm{C}}$, let $C_{\mu,\nu}$ be defined as*

$$C_{\mu,\nu} \coloneqq \max_{h,i,x} \sum_{(\bar{h},\bar{i}) \succeq (h,i)} \mathbb{E}^{x}_{\mu}\left[\left| \sum_{x' \in \mathcal{X}} (T^{\mu}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) - \widehat{T}^{\mu}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}})) V^{\nu,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right|\right].$$

Suppose that $\max_{h,i,x} \|\widehat{T}^{\mu}_{h,i}(\cdot \mid x) - T^{\mu}_{h,i}(\cdot \mid x)\|_1 \leq (4KH)^{-1}$ holds for all $\mu \in \{\mu^{\star}_{\kappa}, \hat{\mu}\}$. Then, we have that:

$$\max_{h,i} \|V^{\star}_{h,i} - V^{\hat{\mu}}_{h,i}\|_\infty \leq 4C_{\mu^{\star}_{\kappa},\mu^{\star}_{\kappa}} + 4C_{\hat{\mu},\mu^{\star}_{\kappa}} + 3H\kappa.$$

Proof. Let us start from Lemma 2 and analyze $V^{\mu^\star_\kappa,\kappa} - V^{\hat{\mu},\kappa}$. For any (h,i) it holds that:

$$\begin{aligned}
\max_{h,i} \|V^{\mu^\star_\kappa,\kappa}_{h,i} - V^{\hat{\mu},\kappa}_{h,i}\|_\infty &\leq \|V^{\mu^\star_\kappa,\kappa}_{h,i} - \widehat{V}^{\mu^\star_\kappa,\kappa}_{h,i}\|_\infty + \|\widehat{V}^{\hat{\mu},\kappa}_{h,i} - V^{\hat{\mu},\kappa}_{h,i}\|_\infty \\
&\leq 2C_{\mu^\star_\kappa,\mu^\star_\kappa} + \|\widehat{V}^{\hat{\mu},\kappa}_{h,i} - V^{\hat{\mu},\kappa}_{h,i}\|_\infty && \text{(Lemma 3 with } \mu = \mu^\star_\kappa) \\
&\leq 2C_{\mu^\star_\kappa,\mu^\star_\kappa} + 2C_{\hat{\mu},\mu^\star_\kappa} + 2KHD_{\hat{\mu},\mu^\star_\kappa} \\
&&& \text{(Lemma 4 with } \mu = \hat{\mu} \text{ and } \nu = \mu^\star_\kappa) \\
&= 2C_{\mu^\star_\kappa,\mu^\star_\kappa} + 2C_{\hat{\mu},\mu^\star_\kappa} + 2KH \max_{\bar{h},\bar{i},\bar{x}} \|\widehat{T}^{\hat{\mu}}_{\bar{h},\bar{i}}(\cdot \mid \bar{x}) - T^{\hat{\mu}}_{\bar{h},\bar{i}}(\cdot \mid \bar{x})\|_1 \max_{\bar{h},\bar{i}} \|V^{\hat{\mu},\kappa}_{\bar{h},\bar{i}} - V^{\mu^\star_\kappa,\kappa}_{\bar{h},\bar{i}}\|_\infty \\
&\leq 2C_{\mu^\star_\kappa,\mu^\star_\kappa} + 2C_{\hat{\mu},\mu^\star_\kappa} + \frac{1}{2} \max_{\bar{h},\bar{i}} \|V^{\hat{\mu},\kappa}_{\bar{h},\bar{i}} - V^{\mu^\star_\kappa,\kappa}_{\bar{h},\bar{i}}\|_\infty, \\
&&& \text{(By assumption on } \widehat{T}^{\hat{\mu}})
\end{aligned}$$

where the equality step follows from the definition of $D_{\hat{\mu},\mu^\star_\kappa}$. Rearranging the terms leads to:

$$\max_{h,i} \|V^{\mu^\star_\kappa,\kappa}_{h,i} - V^{\hat{\mu},\kappa}_{h,i}\|_\infty \leq 4C_{\mu^\star_\kappa,\mu^\star_\kappa} + 4C_{\hat{\mu},\mu^\star_\kappa}.$$

Combining these results with Lemma 2 yields the result. □

C.2.2 High Probability Error Bound

Next, we prove an important auxiliary step that will be subsequently used to upper bound the terms that appeared in Lemma 5.

Lemma 6 (Variance Lemma)**.** *For any policy $\mu \in \Pi_{\mathrm{C}}$, any $(h,i) \in [H] \times [K]$ and any $x \in \mathcal{X}$, it holds that:*

$$\sum_{(\bar{h},\bar{i}) \succeq (h,i)} \mathbb{E}^x_\mu \left[\mathbb{V}\mathrm{ar}_{x' \sim T^\mu_{\bar{h},\bar{i}}(\cdot \mid X_{\bar{h},\bar{i}})} \left(V^{\mu^\star_\kappa,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right) \right] \leq 3H^2.$$

Proof. In the rest of this proof, in order to simplify the notation, we denote $V^{\mu^\star_\kappa,\kappa}_{h,i}$ by $V_{h,i}$. We start by decomposing the variance. Observe that, for all $y \in \mathcal{X}$ and any $(h,i) \in [H] \times [K]$ it holds that:

$$\mathbb{V}\mathrm{ar}_{x' \sim T^\mu_{h,i}(\cdot \mid y)} \left(V_{h^+,i^+}(x') \right) = \sum_{x' \in \mathcal{X}} T^\mu_{h,i}(x' \mid y) V_{h^+,i^+}(x')^2 - \left(\sum_{x' \in \mathcal{X}} T^\mu_{h,i}(x' \mid y) V_{h^+,i^+}(x') \right)^2.$$

Therefore, we can rewrite $\sum_{(\bar{h},\bar{i}) \succeq (h,i)} \mathbb{E}^x_\mu \left[\mathbb{V}\mathrm{ar}_{x' \sim T^\mu_{\bar{h},\bar{i}}(\cdot \mid X_{\bar{h},\bar{i}})} \left(V_{\bar{h}^+,\bar{i}^+}(x') \right) \right]$ as:

$$\sum_{(\bar{h},\bar{i}) \succeq (h,i)} \mathbb{E}^x_\mu \left[\sum_{x' \in \mathcal{X}} T^\mu_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) V_{\bar{h}^+,\bar{i}^+}(x')^2 - \left(\sum_{x' \in \mathcal{X}} T^\mu_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) V_{\bar{h}^+,\bar{i}^+}(x') \right)^2 \right].$$

Let us add and subtract $V_{\bar{h},\bar{i}}(X_{\bar{h},\bar{i}})^2$ within the inner sums over states. Then, we can decompose this equation into the following two terms:

$$\begin{aligned}
A &\coloneqq \sum_{(\bar{h},\bar{i}) \succeq (h,i)} \mathbb{E}^x_\mu \left[\sum_{x' \in \mathcal{X}} T^\mu_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) V_{\bar{h}^+,\bar{i}^+}(x')^2 - V_{\bar{h},\bar{i}}(X_{\bar{h},\bar{i}})^2 \right] \\
B &\coloneqq \sum_{(\bar{h},\bar{i}) \succeq (h,i)} \mathbb{E}^x_\mu \left[V_{\bar{h},\bar{i}}(X_{\bar{h},\bar{i}})^2 - \left(\sum_{x' \in \mathcal{X}} T^\mu_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) V_{\bar{h}^+,\bar{i}^+}(x') \right)^2 \right].
\end{aligned}$$

We first upper bound A. It holds that:

$$\begin{aligned}
A &= \sum_{(\bar{h},\bar{i})\succeq(h,i)} \mathbb{E}_\mu^x \left[\sum_{x'\in\mathcal{X}} T^\mu_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) V_{\bar{h}^+,\bar{i}^+}(x')^2 - V_{\bar{h},\bar{i}}(X_{\bar{h},\bar{i}})^2 \right] \\
&= \sum_{(\bar{h},\bar{i})\succeq(h,i)} \mathbb{E}_\mu^x \left[V_{\bar{h}^+,\bar{i}^+}(X_{\bar{h}^+,\bar{i}^+})^2 - V_{\bar{h},\bar{i}}(X_{\bar{h},\bar{i}})^2 \right] \\
&= \mathbb{E}_\mu^x \left[V_{H,K}(X_{H,K})^2 - V_{h,i}(X_{h,i})^2 \right] \\
&\leq H^2.
\end{aligned}$$

Next, we focus on the upper bound for B. It holds that:

$$\begin{aligned}
B &= \sum_{(\bar{h},\bar{i})\succeq(h,i)} \mathbb{E}_\mu^x \left[V_{\bar{h},\bar{i}}(X_{\bar{h},\bar{i}})^2 - \left(\sum_{x'\in\mathcal{X}} T^\mu_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) V_{\bar{h}^+,\bar{i}^+}(x') \right)^2 \right] \\
&\leq 2H \sum_{(\bar{h},\bar{i})\succeq(h,i)} \mathbb{E}_\mu^x \left[\left(V_{\bar{h},\bar{i}}(X_{\bar{h},\bar{i}}) - \sum_{x'\in\mathcal{X}} T^\mu_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) V_{\bar{h}^+,\bar{i}^+}(x') \right) \right] \\
&= 2H \sum_{(\bar{h},\bar{i})\succeq(h,i)} \mathbb{E}_\mu^x \left[\left(V_{\bar{h},\bar{i}}(X_{\bar{h},\bar{i}}) - V_{\bar{h}^+,\bar{i}^+}(X_{\bar{h}^+,\bar{i}^+}) \right) \right] \\
&= 2H\, \mathbb{E}_\mu^x \left[V_{h,i}(X_{h,i}) - V_{H,K}(X_{H,K}) \right] \\
&\leq 2H^2.
\end{aligned}$$

Here, in the first inequality, we have used that $\mathbb{E}[a^2 - b^2] = \mathbb{E}[(a+b)(a-b)] \leq 2H\mathbb{E}[(a-b)]$ with $a, b \in [0, H]$, and $\mathbb{E}[a] \geq \mathbb{E}[b]$. To see that $\mathbb{E}[a] > \mathbb{E}[b]$, recall that:

$$\begin{aligned}
\mathbb{E}_\mu^x[V_{\bar{h},\bar{i}}(X_{\bar{h},\bar{i}})] &= \mathbb{E}_\mu^x \left[\max_{a\in\mathcal{A}} \sum_{x'\in\mathcal{X}} T_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}, a) V_{\bar{h}^+,\bar{i}^+}(x') \right] \\
&\geq \mathbb{E}_\mu^x \left[\sum_{x'\in\mathcal{X}} T^\mu_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) V_{\bar{h}^+,\bar{i}^+}(x') \right],
\end{aligned}$$

where the first step follows from the the fact that $V_{h,i}$ is the value function of the optimal policy in $\mathcal{M}^{\text{ext}}_\kappa$, *i.e.*, $V_{h,i} = V^{\mu^\star_\kappa,\kappa}_{h,i}$. Combining these results, we obtained that:

$$\sum_{(\bar{h},\bar{i})\succeq(h,i)} \mathbb{E}_\mu^x \left[\mathbb{V}\mathrm{ar}_{x'\sim T^\mu_{\bar{h},\bar{i}}(\cdot|X_{\bar{h},\bar{i}})} \left(V^{\mu^\star_\kappa,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right) \right] = A + B \leq 3H^2,$$

which concludes the proof. □

The following lemma provides a high-probability upper bound on the error of algorithm that we proposed.

Lemma 7 (High Probability Error Bound)**.** *Suppose that m satisfies*

$$m \geq 2 \left(2 + 128K^2H^2 \log\left(\frac{2SAH}{\delta} \right) + 128\, SK^2H^2 \log\left(384\, SK^2H^2 \right) \right). \tag{14}$$

With probability at least $1-\delta$, it holds that:

$$\max_{h,i} \|V^\star_{h,i} - V^{\hat{\mu}}_{h,i}\|_\infty \leq 8\sqrt{\frac{2\log\left(\frac{4|\mathcal{X}|HKA}{\delta}\right) KH^3}{m}} + 8\frac{\log\left(\frac{4|\mathcal{X}|HKA}{\delta}\right) KH^2}{m} + 3H\kappa.$$

Proof. We start by defining a good event $\mathcal{E}$ and prove that it holds with high-probability. Next, we will use $\mathcal{E}$ to prove the result of this lemma. Precisely, we define $\mathcal{E}$ as $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2$, where

$$\mathcal{E}_1 := \bigcap_{(s,a,h)} \left\{ \|\hat{p}_h(\cdot \mid s,a) - p_h(\cdot \mid s,a)\|_1 \leq \sqrt{\frac{2\log\left(\frac{2SAH}{\delta}\right) + 2S\log(6m)}{m}} \right\}$$

$$\mathcal{E}_2 := \bigcap_{(x,a,h,i)} \left\{ \left| \sum_{x' \in \mathcal{X}} \left(T_{h,i}(x' \mid x,a) - \widehat{T}_{h,i}(x' \mid x,a)\right) V^{\mu^\star_\kappa,\kappa}_{h^+,i^+}(x') \right| \right.$$
$$\left. \leq \sqrt{\frac{2\log\left(\frac{4|\mathcal{X}|HKA}{\delta}\right) \mathbb{V}\mathrm{ar}_{x' \sim T_{h,i}(\cdot|x,a)}\left(V^{\mu^\star_\kappa,\kappa}_{h^+,i^+}(x')\right)}{m}} + \frac{2H\log\left(\frac{4|\mathcal{X}|HKA}{\delta}\right)}{3m} \right\}.$$

Then, we first prove that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$. To this end, observe that $\mathbb{P}(\mathcal{E}) = 1 - \mathbb{P}(\mathcal{E}^{\complement}) \geq 1 - \mathbb{P}(\mathcal{E}_1^{\complement}) - \mathbb{P}(\mathcal{E}_2^{\complement})$. Hence, we now upper bound $\mathbb{P}(\mathcal{E}_1^{\complement})$ and $\mathbb{P}(\mathcal{E}_2^{\complement})$. For $\mathbb{P}(\mathcal{E}_1^{\complement})$ and $\mathbb{P}(\mathcal{E}_2^{\complement})$, we use Lemma 9 and Lemma 8 respectively, and obtain

$$\mathbb{P}(\mathcal{E}_1^{\complement}) \leq \sum_{s,a,h} \frac{\delta}{2SAH} \leq \frac{\delta}{2}, \qquad \mathbb{P}(\mathcal{E}_2^{\complement}) \leq \sum_{x,a,h} \frac{\delta}{2|\mathcal{X}|HKA} \leq \frac{\delta}{2}.$$

Therefore, we proved that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$. Next, we show an important property that holds on $\mathcal{E}$. Precisely, for any $\mu \in \Pi_{\mathrm{C}}$, state $x \in \mathcal{X}$ and any (h,i), we have that:

$$\begin{aligned}
\|T^{\mu}_{h,i}(\cdot \mid x) - \widehat{T}^{\mu}_{h,i}(\cdot \mid x)\|_1 &= \sum_{x' \in \mathcal{X}} \left| \sum_{a \in \mathcal{A}} \mu_{h,i}(a \mid x) \left(T_{h,i}(x' \mid x,a) - \widehat{T}_{h,i}(x' \mid x,a)\right) \right| \\
&\leq \sum_{x' \in \mathcal{X}} \sum_{a \in \mathcal{A}} \mu_{h,i}(a \mid x) \left| \left(T_{h,i}(x' \mid x,a) - \widehat{T}_{h,i}(x' \mid x,a)\right) \right| \\
&\leq \max_{a \in \mathcal{A}} \|T_{h,i}(\cdot \mid x,a) - \widehat{T}_{h,i}(\cdot \mid x,a)\|_1 \\
&= \max_{s,a,h} \|p_h(\cdot \mid s,a) - \hat{p}_h(\cdot \mid s,a)\|_1 && \text{(Def. of } T\text{)} \\
&\leq \sqrt{\frac{2\log\left(\frac{2SAH}{\delta}\right) + 2S\log(6m)}{m}} && \text{(Def of } \mathcal{E}_1\text{)} \\
&\leq \frac{1}{4KH}. && \text{(Equation (14) and Lemma 10)}
\end{aligned}$$

Therefore, we proved that:

$$\max_{\mu \in \Pi_{\mathrm{C}}} \max_{h,i} \|T^{\mu}_{h,i}(\cdot \mid x) - \widehat{T}^{\mu}_{h,i}(\cdot \mid x)\|_1 \leq \frac{1}{4KH}. \tag{15}$$

Then, since Equation (15) holds, we can apply Lemma 5, and, with probability at least $1-\delta$ we have that:

$$\max_{h,i} \|V^{\star}_{h,i} - V^{\hat{\mu}}_{h,i}\|_\infty \leq 4C_{\mu^\star_\kappa,\mu^\star_\kappa} + 4C_{\hat{\mu},\mu^\star_\kappa} + 3H\kappa. \tag{16}$$

In the following, we upper bound $C_{\mu^\star_\kappa,\mu^\star_\kappa}$ and $C_{\hat{\mu},\mu^\star_\kappa}$. Before that, for the sake of readability, we introduce some additional notation to simplify the exposition. First, for any pair of policies μ, ν, we denote by $C^{h,i,x}_{\mu,\nu}$ the following quantity:

$$C^{h,i,x}_{\mu,\nu} := \sum_{(\bar{h},\bar{i}) \succeq (h,i)} \mathbb{E}^{x}_{\mu} \left[\left| \sum_{x' \in \mathcal{X}} (T^{\mu}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) - \widehat{T}^{\mu}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}})) V^{\nu,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right| \right],$$

so that $C_{\mu,\nu} = \max_{h,i,x} C^{h,i,x}_{\mu,\nu}$. Furthermore, let $L := \log\left(\frac{4|\mathcal{X}|HKA}{\delta}\right) m^{-1}$. Now, we can start to analyze $C_{\mu,\mu^\star_\kappa}$ for a generic deterministic policy $\mu \in \Pi_{\mathrm{C}}$. Fix a triplet (h,i,x). We have that:

$$
\begin{aligned}
C^{h,i,x}_{\mu,\mu^\star_\kappa} &= \sum_{(\bar{h},\bar{i})\succeq(h,i)} \mathbb{E}^x_\mu \left[\left| \sum_{x'\in\mathcal{X}} (T^\mu_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}) - \widehat{T}^\mu_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}})) V^{\mu^\star_\kappa,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right| \right] \\
&\leq \sum_{(\bar{h},\bar{i})\succeq(h,i)} \mathbb{E}^x_\mu \left[\sum_{a\in\mathcal{A}} \mu(a \mid X_{\bar{h},\bar{i}}) \left| \sum_{x'\in\mathcal{X}} (T_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}, a) - \widehat{T}_{\bar{h},\bar{i}}(x' \mid X_{\bar{h},\bar{i}}, a)) V^{\mu^\star_\kappa,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right| \right] \\
&\qquad \text{(By } |\cdot|\text{)} \\
&\leq \sum_{(\bar{h},\bar{i})\succeq(h,i)} \mathbb{E}^x_\mu \left[\sum_{a\in\mathcal{A}} \mu(a \mid X_{\bar{h},\bar{i}}) \sqrt{2L \mathbb{V}\mathrm{ar}_{x'\sim T_{\bar{h},\bar{i}}(\cdot|X_{\bar{h},\bar{i}},a)} \left(V^{\mu^\star_\kappa,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right)} + LH \right] \quad \text{(By } \mathcal{E}_2\text{)} \\
&\leq \sqrt{2L} \sum_{(\bar{h},\bar{i})\succeq(h,i)} \mathbb{E}^x_\mu \left[\sqrt{\sum_{a\in\mathcal{A}} \mu(a \mid X_{\bar{h},\bar{i}}) \mathbb{V}\mathrm{ar}_{x'\sim T_{\bar{h},\bar{i}}(\cdot|X_{\bar{h},\bar{i}},a)} \left(V^{\mu^\star_\kappa,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right)} \right] + LKH^2 \\
&\qquad \text{(Concavity of } \sqrt{\cdot}\text{)} \\
&\leq \sqrt{2L} \sum_{(\bar{h},\bar{i})\succeq(h,i)} \mathbb{E}^x_\mu \left[\sqrt{\mathbb{V}\mathrm{ar}_{x'\sim T_{\bar{h},\bar{i}}(\cdot|X_{\bar{h},\bar{i}},a)} \left(V^{\mu^\star_\kappa,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right)} \right] + LKH^2 \\
&\leq \sqrt{2L} \sqrt{KH \sum_{(\bar{h},\bar{i})\succeq(h,i)} \mathbb{E}^x_\mu \left[\mathbb{V}\mathrm{ar}_{x'\sim T_{\bar{h},\bar{i}}(\cdot|X_{\bar{h},\bar{i}},a)} \left(V^{\mu^\star_\kappa,\kappa}_{\bar{h}^+,\bar{i}^+}(x') \right) \right]} + LKH^2 \\
&\qquad \text{(Cauchy-Schwarz and Jensen)} \\
&\leq \sqrt{2LKH^3} + LKH^2. \qquad \text{(Lemma 6)}
\end{aligned}
$$

Plugging this result within Equation (16), we have that:

$$\max_{h,i} \|V^\star_{h,i} - V^{\hat{\mu}}_{h,i}\|_\infty \leq 8\sqrt{2LKH^3} + 8LKH^2 + 3H\kappa,$$

which concludes the proof. □

C.2.3 Putting Everything Together

Now, we can finally use Lemma 7 to prove Theorem 4.

Proof of Theorem 4. The proof follows by combining Lemma 7 with an appropriate choice of m and κ. We first $\kappa = \frac{\epsilon}{9H}$. Then, we observe that

$$|\mathcal{X}| \leq S\left(1 + \frac{H}{\kappa}\right)^2 = S\left(1 + \frac{9H^2}{\epsilon}\right)^2.$$

Let

$$L' := \log\left(\frac{4HKAS(1+9H^2/\epsilon)^2}{\delta}\right).$$

Then, we set m as follows:

$$
\begin{aligned}
m &= \max\left\{ 2\left(2 + 128K^2H^2 \log\left(\frac{2SAH}{\delta}\right) + 128\,SK^2H^2 \log\left(384\,SK^2H^2\right)\right), \frac{1152L'KH^3}{\epsilon^2} \right\} \\
&\leq \widetilde{\mathcal{O}}\left(SK^2H^2 \log\left(\frac{1}{\delta}\right) + \frac{KH^3 \log(\frac{1}{\delta})}{\epsilon^2} \right).
\end{aligned}
$$

Now we can apply Lemma 7 and obtain that, with probability at least $1-\delta$:

$$\max_{h,i} \|V^\star_{h,i} - V^{\hat{\mu}}_{h,i}\|_\infty \leq 8\sqrt{\frac{2\log\left(\frac{4|\mathcal{X}|HKA}{\delta}\right) KH^3}{m}} + 8\frac{\log\left(\frac{4|\mathcal{X}|HKA}{\delta}\right) KH^2}{m} + 3H\kappa \leq \epsilon,$$

thus concluding the proof. □

D AUXILIARY LEMMAS

Lemma 8 (Application of Bernstein's Inequality)**.** *Let $D \in \mathbb{N}_{>0}$ and $q \in \Delta(q)$. Let $\hat{q}$ be the maximum likelihood estimator of q computed using n independent samples from q. Furthermore, let $L > 0$ and $f : D \to [0, L]$ be any bounded function. Then, fix $\delta \in (0, 1)$. With probability at least $1 - \delta$ it holds that:*

$$|(q - \hat{q})^\top f| \leq \sqrt{\frac{2 \log\left(\frac{2}{\delta}\right) \mathbb{V}\mathrm{ar}_q(f)}{n}} + \frac{2L \log\left(\frac{2}{\delta}\right)}{3n}.$$

Proof. Direct application of Bernstein's inequality (Boucheron et al., 2003). □

Lemma 9 (Concentration of $\|\cdot\|_1$ norms)**.** *Consider $D \in \mathbb{N}_{>0}$ and $q \in \Delta(D)$. Let $\hat{q}$ be the maximum likelihood estimator of q after n i.i.d. samples. Then, for $\delta \in (0, 1)$, it holds that:*

$$\mathbb{P}\left(\|q - \hat{q}\|_1 \geq \sqrt{\frac{2 \log\left(\frac{1}{\delta}\right) + 2D \log(6n)}{n}}\right) \leq \delta.$$

Proof. Direct application of Pinsker's inequality and Proposition 1 in Jonsson et al. (2020). □

Lemma 10 (Technical Lemma; Lemma D.9 in Barbara et al. (2026))**.** *Let $c_1, c_2, n, L, \delta \in \mathbb{R}_{>0}$ and consider*

$$\bar{T} = \inf\left\{t \in \mathbb{N} : \frac{\log(c_1/\delta) + n \log(c_2 t)}{t} \leq L^2\right\}.$$

It holds that:

$$\bar{T} \leq 2\left(2 + \frac{4 \log(c_1/\delta)}{L^2} + \frac{4n}{L^2} \log\left(\frac{2nc_2}{L^2}\right)\right).$$

Lemma 11 (Lemma 1 in Garivier et al. (2019))**.** *Consider a measurable space $(\Omega, \mathcal{F})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$. For any $\mathcal{F}$-measurable function $Z : \Omega \to [0, 1]$ we have:*

$$\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2) \geq \mathrm{kl}(\mathbb{E}_1[Z], \mathbb{E}_2[Z]),$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are the expectations under $\mathbb{P}_1$ and $\mathbb{P}_2$ and $\mathrm{kl}(p, q)$ denotes the KL-divergence between Bernoulli distributions with mean p and q.

Lemma 12 (Lemma 15 in Domingues et al. (2021))**.** *For any $p, q \in [0, 1]$:*

$$\mathrm{kl}(p, q) \geq (1 - p) \log\left(\frac{1}{1 - q}\right) - \log(2),$$

where $\mathrm{kl}(p, q)$ denotes the KL-divergence between Bernoulli distributions with mean p and q.