QUANTILE OF MEANS: A BONUS-FREE ENSEMBLE METHOD FOR MINIMAX OPTIMAL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimal Reinforcement Learning (RL) algorithms typically rely on carefully constructed count-based uncertainty estimates, such as bonuses or posterior distributions, to drive exploration. Although theoretically sound, such estimates are difficult to scale and therefore offer limited insight for designing exploration heuristics in practice. Meanwhile, ensembling has emerged as a practical approach, but remains without theoretical justification. Building on a recent ensemble-based method for Multi-Armed Bandits, we propose a quantile-based ensemble method for finite-horizon tabular MDPs. Our simple bonus-free approach achieves optimal variance-dependent regret bounds without explicit uncertainty estimation, providing theoretical grounding for ensemble-based exploration in RL.

1 INTRODUCTION

Reinforcement learning (RL) provides a general framework for sequential decision-making under uncertainty, where an agent interacts with an environment to optimize a long-term objective. RL algorithms have achieved remarkable success across domains ranging from robotics and games (Mnih et al., 2015; Schulman et al., 2015; 2017; Haarnoja et al., 2018) to Large Language Models (LLMs; Stiennon et al. (2020); Ouyang et al. (2022)). Despite this progress, designing algorithms that can efficiently explore their environment – gathering informative experience while avoiding excessive trial and error – remains one of the fundamental challenges in RL.

To balance exploration and exploitation, optimal algorithms traditionally rely on explicitly quantifying uncertainty—through confidence bonuses or posterior distributions—to bias their value estimates optimistically, effectively inflating the value of unknown actions to encourage exploration. However, in environments requiring function approximation, explicitly calculating these bounds often yields estimates that are either computationally intractable or vacuous for realistic sample sizes. This highlights a fundamental bottleneck of explicit methods: their algorithmic bias increases proportionally to how tightly they can quantify uncertainty. Because these explicit uncertainty estimates break down in high-dimensional domains, such methods provide limited guidance for designing scalable, practical exploration heuristics.

In contrast, ensemble-based methods have become a staple of modern deep reinforcement learning (Lee et al., 2021; Chen et al., 2021; Osband et al., 2016a;b; Pathak et al., 2019; Peer et al., 2021). These methods attempt to drive exploration by leveraging the natural variation across a collection of standard approximators. Because they do not rely on explicit uncertainty bounds, their added bias is dictated entirely by the ensemble size and the aggregation strategy used to combine predictions. However, the theoretical principles governing these choices remain largely unexplored, even in the foundational tabular Markov Decision Process (MDP) setting. For instance, while taking a minimum across the ensembles might intuitively encourage exploration, this remained unproven.

Recent progress has begun to bridge this gap. In Cassel et al. (2025), a pure batch ensemble scheme relying precisely on this minimum aggregation was shown to achieve near-optimal, variance-dependent regret bounds for stochastic Multi-Armed Bandits (MAB). Viel et al. (2025) recently attempted to adapt a similar ensemble approach to MDPs in an imitation learning context, their

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.

method achieved suboptimal rates and required binarizing the state space—a construction that scales quadratically and cannot be applied heuristically in a function approximation setting.

Consequently, several fundamental questions regarding the viability of pure ensemble exploration for sequential decision-making remain unresolved: (1) Can pure ensemble exploration yield non-trivial regret in MDPs without relying on explicit uncertainty quantification? (2) If so, does a small ensemble—one scaling logarithmically with the problem parameters—suffice? (3) And finally, is it possible to aggregate ensemble predictions without inherently leaving performance on the table? Resolving these questions in the tabular setting is an essential first step toward demystifying ensemble exploration.

Contributions. In this work, we answer these fundamental questions affirmatively, establishing the first rigorous foundation for pure ensemble exploration in MDPs. Our algorithm, **Value Iteration with Bootstrap Ensemble (VIBE)**, selects actions based on a fixed quantile of independent Q-value estimates. By employing this Quantile of Means (QoM) aggregation strategy, we achieve minimax optimal regret in tabular MDPs without ever relying on explicit exploration bonuses, posterior computations, or prior knowledge of reward distributions. Crucially, this optimal performance is guaranteed using a small ensemble size that scales logarithmically with the problem parameters—making the algorithm computationally viable and providing a theoretical justification for the small ensemble sizes used in deep RL practice. Ultimately, our results demonstrate that ensembles provide a theoretically sound—yet scalable—alternative to explicit uncertainty quantification.

Beyond this conceptual foundation, our QoM framework yields several additional benefits:

- **Distribution-agnostic algorithm, distribution-adaptive performance.** Our algorithm encodes no distributional assumptions, functioning seamlessly across (non-negative) bounded, sub-Gaussian, and heavy-tailed loss distributions. Because we bypass explicit bonuses, the regret naturally adapts to the true concentration properties of the underlying distributions.
- **Variance-dependent regret.** We prove that *VIBE* achieves optimal variance-dependent regret, matching the best known results (Zanette & Brunskill, 2019; Zhou et al., 2023)—previously attainable only through complex count-based bonus constructions.
- **Optimal bandit ensembles.** By applying our QoM estimator to the batch ensemble framework of Cassel et al. (2025), we shave a logarithmic factor from their regret, achieving the optimal instance-dependent rate (see Section C).
- **Simple and transparent analysis.** The absence of explicit bonus terms makes the key arguments considerably more transparent than in prior variance-dependent analyses, allowing us to present a (nearly) complete proof in the main text.
- **Empirical proof of concept.** We provide empirical evaluations to bridge our theoretical findings with practice (see Section D). The results confirm that *VIBE* is competitive out-of-the-box and demonstrate its potential to surpass standard methods when ensemble sizes are scaled according to a tighter constant conjectured by Feige (2004).

Extending ensemble-based exploration from bandits to MDPs is substantially more challenging than a per-state reduction: Q-value estimates at each step depend on subsequent value estimates through Bellman backups, breaking the independence structure that makes the bandit analysis clean. Standard covering arguments over the value function address this dependence but result in an ensemble size linear in the number of states, rather than logarithmic. Achieving optimal dependence requires advanced analytical techniques that, in prior work, are always intertwined with intricate bonus constructions (Zanette & Brunskill, 2019; Zhou et al., 2023). Whether our bonus-free approach extends to linear or general function approximation settings remains an interesting open question.

1.1 Related work

Variance-dependent regret in MDPs. Achieving minimax optimal regret in tabular MDPs has been the focus of a long line of work (Azar et al., 2017; Dann et al., 2017; Jin et al., 2018), with tight bounds scaling as $\tilde{O}(\sqrt{H^3SAT})$. Building on this, Zanette & Brunskill (2019) provided the first variance-dependent regret bounds, introducing carefully constructed count-based bonuses that yield a highly complex algorithm and analysis. Zhou et al. (2023) recently refined this variance dependence further and established matching lower bounds, confirming the tightness of such problem-dependent results. However, both methods fundamentally require explicit uncertainty quantification

through bonuses. In contrast, our algorithm achieves the same variance-dependent guarantees using a simple quantile-based ensemble mechanism—requiring no bonuses, explicit variance estimation, or prior knowledge of reward distributions—while matching the lower bounds of Zhou et al. (2023).

Posterior/quantile-based exploration in MDPs. Our work also conceptually relates to Posterior Sampling for RL (PSRL) and randomized value functions (Osband et al., 2013; Agrawal & Jia, 2017; Russo, 2019). More recently, Tiapkin et al. (2022b) proposed Bayes-UCBVI, which uses quantiles of a posterior over Q-values to achieve optimal $\tilde{O}(\sqrt{H^3SAT})$ regret without explicit bonuses. Similarly, Tiapkin et al. (2022a) take an ensembling approach using a few posterior samples. Both algorithms, however, fundamentally rely on maintaining a Dirichlet posterior over transition probabilities—an inherently count-based statistical operation—and assume the reward function is fully known. Adapting these methods to unknown reward distributions requires strong prior knowledge. Our method achieves comparable optimism using a pure data-partitioning ensemble that requires no explicit posterior computation or distributional assumptions.

Ensemble methods in bandits. Our work directly builds on Cassel et al. (2025), which introduced the batch ensemble scheme for stochastic bandits achieving variance-dependent regret without distributional assumptions. This paper is the first to extend this strictly bonus-free approach to MDPs. Other bandit perturbation and ensemble techniques (Lu & Van Roy, 2017; Kveton et al., 2019; Lee & Oh, 2024) or bootstrap bounds (Hao et al., 2019) either fail to adapt to unknown reward distributions, inject artificial noise, or strictly require symmetric rewards. Furthermore, while Bayesian approaches like Bayes-UCB (Kaufmann et al., 2012) use posterior quantiles, they still fundamentally rely on computing explicit count-dependent confidence levels.

Practical ensemble methods in deep RL. Ensembles are a key technique in empirical deep RL exploration. Methods such as Bootstrapped DQN (Osband et al., 2016a;b), UCB Q-ensembles (Chen et al., 2017), Ensemble Bootstrapped Q-Learning (Peer et al., 2021), and SUNRISE (Lee et al., 2021) have demonstrated remarkable empirical success in complex environments. However, the theoretical justification for why these ensemble heuristics successfully induce exploration has remained largely open. Our work bridges this gap, providing a rigorous theoretical foundation for bonus-free ensemble-driven exploration in RL.

2 The Quantile of Means (QoM) Estimator

To build the foundation for our ensemble-based MDP algorithm, we first isolate the core statistical challenge: estimating the mean of a distribution optimistically from a finite sample without relying on explicit uncertainty bonuses. To this end, we introduce the Quantile of Means (QoM) estimator, a generalization of the classic Median of Means (MoM) approach.

Let $X_1, \dots, X_n$ be i.i.d *nonnegative* random variables, with expectation μ. First, for any sequence $\hat{\mu}^1, \dots, \hat{\mu}^B \in \mathbb{R}$ define its quantile at level $\alpha \in [0, 1]$ as

$$q_\alpha(\hat{\mu}^b, b \in [B]) \stackrel{\text{def}}{=} \hat{\mu}^{(\lceil \alpha B \rceil)}, \tag{1}$$

where $\hat{\mu}^{(1)}, \dots, \hat{\mu}^{(B)}$ is the sequence sorted in ascending order, and $\lceil \cdot \rceil$ is the ceiling function (which rounds numbers towards infinity). Then, given a partition of the n samples into B fixed, disjoint subsets $\mathcal{D}^1, \dots, \mathcal{D}^B$, the Quantile of Means estimator is defined as:

$$\hat{\mu}_\alpha \stackrel{\text{def}}{=} q_\alpha \left(\sum_{X \in \mathcal{D}^b} \frac{X}{|\mathcal{D}^b| + 1}, b \in [B] \right). \tag{2}$$

We establish the following guarantees for the QoM estimator (see proof in Section A.1).

Lemma 1. *Let $X_1, \dots, X_n$ be i.i.d. non-negative random variables bounded almost surely by R, with expectation μ and variance σ^2. Let $\hat{\mu}_\alpha$ be the QoM estimator (Eq. (2)) instantiated with $B \geq 26 \log \delta^{-1}$ batches and quantile level $\alpha = 1/65$. Then, each of the following bounds holds individually with probability at least $1 - \delta$:*

1. *(*Optimism*)* $\hat{\mu}_\alpha \leq \mu$.
2. *(*Bias*) If $|\mathcal{D}^b| \geq \lfloor n/B \rfloor$ for all $b \in [B]$, then: $\hat{\mu}_\alpha \geq \mu - 1.7\sqrt{\frac{\sigma^2 B}{\max\{1,n\}}} - \frac{9RB}{\max\{1,n\}}$.*

Remark (Bernstein-equivalent scaling). To see that the QoM estimator incurs no asymptotic efficiency loss compared to the standard empirical mean, we can substitute the minimal ensemble size $B = \Theta(\log \delta^{-1})$ into the bias bound. This recovers the standard Bernstein concentration scaling:

$$\hat{\mu}_\alpha \geq \mu - O\left(\sqrt{\frac{\sigma^2 \log \delta^{-1}}{\max\{1,n\}}} + \frac{R \log \delta^{-1}}{\max\{1,n\}}\right).$$

Crucially, this demonstrates that the statistical inefficiency introduced by data partitioning does not scale with any problem parameters, but rather acts strictly as a constant factor penalty hidden in the asymptotic notation.

Standard estimators achieve *optimism* by explicitly subtracting an uncertainty estimate from an unbiased mean estimator. Consequently, bounding their bias requires carefully designing and computing tight uncertainty bonuses. In contrast, QoM achieves optimism purely through data partitioning and order statistics. This mechanism relies on the following corollary, which states that a minor modification to the standard sample average is optimistic with constant probability.

Corollary 2. *Let $X_1, \ldots, X_n$ be i.i.d nonnegative random variables with expectation μ. Then for any $c \geq 1/12$ it holds that:* $\Pr\left(\sum_{i=1}^{n} \frac{X_i}{n+c} < \mu\right) \geq 1/13$.

Proof. Apply Lemma 13 in Section A.3 (Feige (2004), Theorem 1) with X_i/μ and $\delta = c$. ■

Crucially, this is not a standard concentration inequality and requires no information beyond the first moment. Using Corollary 2, the QoM estimator is optimistic if at least $\lceil \alpha B \rceil$ of the subsets yield optimistic estimates. Because the B subsets are independent, the number of optimistic subsets follows a Binomial distribution with B trials and a success probability of at least $1/13$. Bounding the lower tail of this Binomial via a Chernoff bound yields the optimism guarantee in Lemma 1.

The *bias* property of QoM follows a similar logic, replacing Corollary 2 with Freedman's inequality (Lemma 14 in Section A.3). Similar guarantees can be obtained using other tail inequalities. For example, applying Chebyshev's inequality allows us to handle unbounded random variables with finite variance, albeit with significantly worse constants. Importantly, adapting to these different distributions requires essentially no changes to the QoM estimator itself—only to the analysis. This stands in stark contrast to bonus-based estimators, where the algorithmic implementation must change to incorporate distribution-specific bonus terms.

Finally, we note that Cassel et al. (2025) recently considered the Minimum of Means estimator, which is equivalent to QoM with $\alpha = 1/B$. They provide similar guarantees in their Lemmas 3 and 4; however, their optimism claim is applicable only for Bernoulli or symmetric random variables, and their bias bound incurs additional logarithmic factors.

3 Setting and Notations

MDP. A finite horizon MDP is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, \mathcal{L})$, where $\mathcal{S}$ is a set of states (of size S), $\mathcal{A}$ is a set of actions (of size A), H is the decision horizon, $P_h, h \in [H]$ are the transition dynamics, and $\mathcal{L}$ is the loss distribution. Each episode $k \in [K]$ starts at state $s_1 \in \mathcal{S}$ and proceeds inductively as follows. If at step $h \in [H]$ the system is in state $s_h^k \in \mathcal{S}$, and an agent chooses action $a_h^k \in \mathcal{A}$, then the system transitions to a new state $s_{h+1}^k \in \mathcal{S}$ sampled independently according to $P_h(\cdot \mid s_h^k, a_h^k)$, and a loss $L_h^k \in \mathbb{R}$ is sampled independently according to $\mathcal{L}_h(s_h^k, a_h^k)$ whose mean is denoted $\ell_h(s_h^k, a_h^k)$. The process ends after H steps, at which point a new episode begins. We note that fixing the initial state is done to ease the notation and does not lose generality.

Assumption 3 (Bounded total loss). *The random losses L_h are non-negative and satisfy $\sum_{h\in[H]} L_h \leq H$ almost surely for all policies.*

Policy and Regret. We consider agent strategies known as (deterministic) Markov policies $\pi = (\pi_h)_{h\in[H]} : [H] \times \mathcal{S} \mapsto \mathcal{A}$, which map a step and state to an action. Such a policy induces a distribution over trajectories $\iota = (s_h, a_h)_{h\in[H]}$, and we denote expectation with respect to this distribution as $\mathbb{E}_{P,\pi}[\cdot]$. For each policy π and horizon $h \in [H]$ we define its value (or loss to-go)

as $V_h^\pi(s) = \mathbb{E}_{P,\pi}[\sum_{h'=h}^{H} \ell_{h'}(s_{h'}, a_{h'}) \mid s_h = s]$, which is the expected loss if one starts from state s at horizon h and follows policy π. The performance of a policy, also called its value, is measured by its expected cumulative loss, given by $V_1^\pi(s_1)$. Thus, the optimal policy and value are given by $\pi^\star \in \arg\min_{\pi \in \Pi_M} V^\pi(s_1)$ and $V^\star = V_1^{\pi^\star}$, where Π_M is the class of deterministic Markov policies, which is known to be optimal even among the class of stochastic history-dependent policies. Finally, we measure the quality of any algorithm via its *regret* – the difference between the value of the policies π^k generated by the algorithm and that of the optimal policy $\pi^\star$, i.e.,

$$\text{regret}_K = \sum_{k \in [K]} V_1^{\pi^k}(s_1) - V_1^\star(s_1).$$

Additional notation. We use the following notations throughout. $[m]$ for any $m \in \mathbb{N}$, denotes the set $\{1, \dots, m\}$. For any $V : \mathcal{S} \to \mathbb{R}, s \in \mathcal{S}, a \in \mathcal{A}$ denote $PV(s,a) = \sum_{s' \in \mathcal{S}} P(s' \mid s,a)V(s')$. For a random variable X, let $\text{Var}_{s,a,h}(X)$ be its variance conditioned on $(s_h, a_h) = (s,a)$.

4 Algorithm and Main Result

We now introduce Value Iteration with Bootstrap Ensemble (VIBE; Algorithm 1), an algorithm that replaces explicit uncertainty bonuses with our QoM estimator. Standard optimistic algorithms construct explicit upper confidence bounds by estimating the empirical mean and subtracting a bonus proportional to the inverse square root of the visitation count. VIBE bypasses this entirely. Instead, we maintain B independent, non-overlapping datasets $\mathcal{D}_h^{k,b}(s,a)$ for each state-action-step tuple.

Concretely, to ensure these datasets grow at the same rate—thereby maintaining a balanced variance across the ensemble—we assign incoming transitions (s_{h+1}^k, L_h^k) using a strict round-robin schedule:

$$\mathcal{D}_h^{k+1,b}(s,a) = \mathcal{D}_h^{k,b}(s,a) \bigcup \begin{cases} (s_{h+1}^k, L_h^k), & \text{if } (s,a,b) = (s_h^k, a_h^k, b_h^k(s,a)) \\ \emptyset, & \text{otherwise,} \end{cases} \tag{3}$$

where $b_h^k(s,a) = \arg\min_{b \in [B]} |\mathcal{D}_h^{k,b}(s,a)|$. The round-robin sample assignment is an analytical device used solely to guarantee exact statistical independence across estimators in the proof. In practical deep RL implementations, no counts are tracked: one simply trains an ensemble of B neural networks on random mini-batches and evaluates the upper quantile of their Q-value predictions.

During planning, we perform standard value iteration (without bonuses), but evaluate the $Q-$value using a QoM estimate (see Eq. (2)) applied across our B (nearly) equal non-overlapping datasets at quantile level α. Because we are minimizing loss, taking the α-quantile of the Q-value estimates naturally filters out overestimations while implicitly establishing a high-probability optimistic lower bound on the true loss-to-go. The full procedure is detailed in Algorithm 1.

From a practical standpoint, VIBE is highly tunable. It requires only two hyperparameters: the quantile level α (which is constant) and the number of batches B (which scales logarithmically with H, S, K, and δ^{-1}). Computationally, VIBE runs value iteration across B ensemble members, taking $O(S^2AHB)$ time per episode sequentially – a factor B overhead. Crucially, because updates are parallelizable across B threads, wall-clock time remains almost identical to standard value iteration.

Main result. To state our theoretical guarantees, we characterize the problem complexity using the maximum cumulative conditional variance of the MDP, similarly to Zanette & Brunskill (2019). We define $\mathbb{Q}^\star = \sum_{h \in [H]} \mathbb{Q}_h^\star$ where

$$\mathbb{Q}_h^\star = \max_{s \in \mathcal{S}, a \in A} \text{Var}_{s,a,h}(L_h + V_{h+1}^\star(s_{h+1})). \tag{4}$$

Theorem 4 establishes that VIBE achieves a variance-dependent regret bound without requiring explicit variance estimation or bonuses. See proof in Section 6, and discussion on extending to heavy-tailed losses in Section B.

Theorem 4. *Suppose we run Algorithm 1 with $\alpha = 1/65$ and $B = 26\log(5SAHK\delta^{-1})$, and let $\kappa = \log(20HS^2AK\delta^{-1})$ be a logarithmic factor. With probability at least $1-\delta$,*

$$\text{regret}_K \le 22\sqrt{\min\{\mathbb{Q}^\star, HV^\star\}HSAK\kappa^2} + 1924H^3S^2A\kappa^3.$$

Algorithm 1 Value Iteration with Bootstrap Ensemble (VIBE)

input: Number of batches B, quantile level α.
initialize: Datasets $\mathcal{D}_h^{1,b}(s,a) = \emptyset$ for all $s \in S, a \in A, h \in [H]$.
for episode $k = 1, 2, \ldots, K$ **do**
  Define $\hat{V}_{H+1}^k(s) = 0$ for all $s \in S$.
  for $h = H, \ldots, 1$ **do**
    for $b \in [B]$ **do**
      $\hat{P}_h^{k,b}(s' \mid s,a) = \sum_{s_+ \in \mathcal{D}_h^{k,b}(s,a)} \frac{\mathbb{1}\{s_+ = s'\}}{|\mathcal{D}_h^{k,b}(s,a)|+1}$
      $\hat{\ell}_h^{k,b}(s,a) = \sum_{L \in \mathcal{D}_h^{k,b}(s,a)} \frac{L}{|\mathcal{D}_h^{k,b}(s,a)|+1}$
      $\hat{Q}_h^{k,b}(s,a) = [\hat{\ell}_h^{k,b} + \hat{P}_h^{k,b}\hat{V}_{h+1}^k](s,a)$
    $\hat{V}_h^k(s) = \min_{a \in \mathcal{A}} q_\alpha\Big(\hat{Q}_h^{k,b}(s,a), b \in [B]\Big)$ ▷ QoM Eq. (2)
  Play π^k defined as $\pi_h^k(s) \in \arg\min_{a \in \mathcal{A}} q_\alpha\Big(\hat{Q}_h^{k,b}(s,a), b \in [B]\Big)$ and observe trajectory ι^k.
  Add each (s_{h+1}^k, L_h^k) to $\mathcal{D}_h^{k+1,b}(s_h^k, a_h^k)$ with fewest samples. ▷ Round-Robin Eq. (3)

Crucially, the leading term of this bound establishes the optimality of our ensemble approach in two distinct regimes (up to logarithmic factors). In the worst case, it matches the standard $\Omega(\sqrt{H^3SAK})$ minimax lower bound of (Domingues et al., 2021). Moreover, in environments with low stochasticity, the bound scales with the problem-dependent variance $\mathcal{Q}^*$, matching the refined variance-dependent lower bounds established by Zhou et al. (2023).

The lower-order burn-in term is independent of the number of episodes K and optimizing it is not the focus of this work. The theoretical batch size B acts as an effective confidence parameter analogous to $\log(1/\delta)$ in standard Bernstein concentration. Furthermore, the constants in our bound stems from Feige (2004), Theorem 1, which guarantees constant probability of $1/13$. Under their conjecture that this constant in fact equals $1/e \approx 0.368$, the required ensemble size B shrinks by a factor of $e/13 \approx 0.209$, reducing both the lower-order regret term and computational overhead proportionally. We empirically validate the practical impact of this conjectured scaling in Section D, demonstrating that it allows VIBE to outperform standard baselines in a tabular environment.

5 Ensembles as a Tool for Optimism

Having established VIBE's theoretical guarantees and algorithmic simplicity, we now dissect the mathematical engine driving these results. The purpose of this section is to formally demonstrate how our ensemble quantile mechanism naturally induces optimism in the MDP setting, bypassing the need for explicit uncertainty bonuses. To see this, we first decompose the cumulative regret into two standard components—colloquially referred to as bias and optimism:

$$\text{regret}_K = \sum_{k \in [K]} \underbrace{V_1^{\pi^k}(s_1) - \hat{V}_1^k(s_1)}_{\text{bias}} + \sum_{k \in [K]} \underbrace{\hat{V}_1^k(s_1) - V_1^\star(s_1)}_{\text{optimism}}. \tag{5}$$

The primary goal of optimistic algorithms is to ensure that the optimism term is non-positive, and thus the total regret is bounded entirely by the bias term. Traditionally, bounding the optimism term requires carefully designed count-based bonuses, which then become the dominant factor when analyzing the bias. Because VIBE completely forgoes explicit bonuses, our bias term analysis is significantly simplified and relies mostly on standard techniques (detailed in Section 6.1). In what follows, we show that our QoM approach achieves optimism on the following high-probability event.

Lemma 5 (Good event for optimism). *Suppose that $\alpha = 1/65$ and $B \geq 26\log(5SHK\delta^{-1})$, then with probability at least $1 - \delta/5$ simultaneously for all $h \in [H], s \in \mathcal{S}, k \in [K]$.*

$$q_\alpha\Big([\hat{\ell}_h^{k,b} + \hat{P}_h^{k,b}V_{h+1}^\star](s, \pi_h^\star(s)), b \in [B]\Big) \leq [\ell_h + P_h V_{h+1}^\star](s, \pi_h^\star(s)).$$

Proof. First, plugging in the definitions of $\hat{\ell}$ and $\hat{P}$ (see Algorithm 1), we have

$$[\hat{\ell}_h^{k,b} + \hat{P}_h^{k,b} V_{h+1}^\star](s,a) = \sum_{L,s' \in \mathcal{D}_h^{k,b}(s,a)} \frac{L + V_{h+1}^\star(s')}{|\mathcal{D}_h^{k,b}(s,a)| + 1}.$$

Moreover, the datasets $\mathcal{D}_h^{k,b}(s,a)$ are disjoint and each contain i.i.d. samples that satisfy $\mathbb{E}[L + V_{h+1}^\star(s')] = [\ell_h + P_h V_{h+1}^\star](s,a)$. Thus, the proof is concluded by performing a union bound on Lemma 1 with respect to $h \in [H], s \in \mathcal{S}$, and all valid configurations of dataset sizes $|\mathcal{D}_h^{k,b}(s, \pi^\star(s))|, b \in [B]$. Because we use a round-robin schedule, there are at most K valid configurations, and thus the total number of events is bounded by SHK, as desired. ■

Lemma 6 (Optimism). *Conditioning on the good event defined in Lemma 5, we have* $\hat{V}_h^k(s) \leq V_h^\star(s)$ *for all* $s \in \mathcal{S}, h \in [H], k \in [K]$.

Proof. We prove the claim by reverse induction on $h \in [H+1]$. The base case is satisfied because $\hat{V}_{H+1}^k = V_{H+1}^\star = 0$. Now, assuming the hypothesis holds for $h+1$, we have

$$\begin{aligned}
\hat{V}_h^k(s) &= \min_{a \in \mathcal{A}} q_\alpha\Big(\hat{Q}_h^{k,b}(s,a), b \in [B]\Big) \\
&\leq q_\alpha\Big(\hat{Q}_h^{k,b}(s, \pi^\star(s)), b \in [B]\Big) \\
&= q_\alpha\Big([\hat{\ell}_h^{k,b} + \hat{P}_h^{k,b} \hat{V}_{h+1}^k](s, \pi_h^\star(s)), b \in [B]\Big) \\
&\leq q_\alpha\Big([\hat{\ell}_h^{k,b} + \hat{P}_h^{k,b} V_{h+1}^\star](s, \pi_h^\star(s)), b \in [B]\Big) && \text{(induction hypothesis)} \\
&\leq [\ell_h + P_h V_{h+1}^\star](s, \pi_h^\star(s)) && \text{(Good event - Lemma 5)} \\
&= V_h^\star(s), && \text{(Bellman optimality equation)}
\end{aligned}$$

where the second and third inequalities also used that $q_\alpha(x) \leq q_\alpha(y)$ for $x \leq y$ (point-wise). ■

6 Regret Analysis

As established in Section 5, our ensemble approach ensures the optimism term in the regret decomposition is non-positive. Consequently, the regret of VIBE is bounded entirely by the bias. To formalize this, we first define our empirical counters and operators. Let $N_h^k(s,a) = \sum_{k' \in [k-1]} \mathbb{1}\Big\{(s,a) = (s_h^{k'}, a_h^{k'})\Big\}$ be the total visitation count. Because the data is distributed across B batches, the effective inverse count per batch is given by $\beta_h^k(s,a) = B/\max\{1, N_h^k(s,a)\}$. Let $\hat{\mathcal{T}}_h^k$ denote the empirical QoM Bellman operator:

$$\hat{\mathcal{T}}_h^k V(s,a) = q_\alpha\Big([\hat{\ell}_h^{k,b} + \hat{P}_h^{k,b} V](s,a), b \in [B]\Big).$$

We define the good event $\mathcal{E}$ as the intersection of several high-probability inequalities that hold simultaneously for all $s, s' \in \mathcal{S}, a \in \mathcal{A}, h \in [H], k \in [K]$, and $b \in [B]$:

$$\hat{\mathcal{T}}_h^k V_{h+1}^\star(s, \pi_h^\star(s)) \leq [\ell_h + P_h V_{h+1}^\star](s, \pi_h^\star(s)) \tag{6}$$

$$\hat{\mathcal{T}}_h^k V_{h+1}^\star(s,a) \geq [\ell_h + P_h V_{h+1}^\star](s,a) - 1.7\sqrt{\beta_h^k(s,a) \operatorname*{Var}_{s,a,h}(L_h + V_{h+1}^\star(s_{h+1}))} - 9H\beta_h^k(s,a) \tag{7}$$

$$\hat{P}_h^{k,b}(s' \mid s,a) \leq \left(1 + \frac{1}{3H}\right) P_h(s' \mid s,a) + 2.2H\kappa\beta_h^k(s,a) \tag{8}$$

$$\hat{P}_h^{k,b}(s' \mid s,a) \geq \left(1 - \frac{1}{3H}\right) P_h(s' \mid s,a) - 2.2H\kappa\beta_h^k(s,a) \tag{9}$$

$$\sum_{k \in [K]} \mathbb{E}_{\pi^k, P} \sum_{h \in [H]} \beta_h^k(s_h, a_h) \leq 2BHSA\kappa. \tag{10}$$

These conditions capture the essential statistical properties needed to bound the regret. Equation equation 6 establishes optimism, ensuring the QoM estimator does not overestimate the loss of the optimal action. Equation equation 7 bounds the algorithm's bias by limiting how much the estimator can underestimate the true Bellman target for any action. Equations equation 8 and equation 9 ensure the empirical transition probabilities concentrate around the true dynamics. Notice that the error here scales multiplicatively with the true transition probabilities, which is a strict requirement for achieving variance-dependent bounds. Finally, Equation equation 10 bounds the cumulative sum of the statistical errors (the inverse counts) encountered by the agent across all episodes.

The following two lemmas establish that $\mathcal{E}$ holds with high probability, and that under $\mathcal{E}$, the cumulative bias is bounded by the problem's structural variance (proofs in Sections A.2 and 6.1).

Lemma 7 (Good Event). *If* $\alpha = 1/65$ *and* $B \geq 26\log(5SAHK\delta^{-1})$ *then* $\Pr(\mathcal{E}) \geq 1-\delta$.

Lemma 8 (Bias). *Suppose that the good event (Eqs.* (6) *to* (10)*) holds, then for any* $\eta > 0$

$$V^{\pi^k}(s_1) - \hat{V}^k(s_1) \leq \frac{\min\{\mathbb{Q}^\star, HV^\star\}}{\eta} + (8.5\eta + 37H^2S\kappa)\mathbb{E}_{P,\pi^k}\left[\sum_{h\in[H]} \beta_h^k(s_h, a_h)\right].$$

With these tools, the proof of our main result (Theorem 4) follows by conditioning on the good event, bounding the optimism term by 0, summing the bias term over all K episodes, using Eq. (10) and optimizing the free parameter $\eta = \sqrt{\min\{\mathbb{Q}^\star, HV^\star\}K/17BHSA\kappa}$.

6.1 Proof of Lemma 8 (Bias Bound)

To ease notation, let $\hat{V}_h^{k,b}(s) = [\hat{\ell}_h^{k,b} + \hat{P}_h^{k,b}\hat{V}_{h+1}^k](s, \pi_h^k(s))$ be the predicted value of batch $b \in [B]$ at step $h \in [H]$ of episode $k \in [K]$. Then, for any $s \in \mathcal{S}, h \in [H], k \in [K], b \in [B]$, we decompose the error using the Bellman equations:

$$\begin{aligned} V_h^{\pi^k}(s) - \hat{V}_h^{k,b}(s) &= [(\ell_h - \hat{\ell}_h^{k,b}) + P_h V_{h+1}^{\pi^k} - \hat{P}_h^{k,b}\hat{V}_{h+1}^k](s, \pi_h^k(s)) \qquad (11)\\ &= [(\ell_h - \hat{\ell}_h^{k,b}) + (P_h - \hat{P}_h^{k,b})V_{h+1}^\star + (P_h - \hat{P}_h^{k,b})(V_{h+1}^{\pi^k} - V_{h+1}^\star) + \hat{P}_h^{k,b}(V_{h+1}^{\pi^k} - \hat{V}_{h+1}^k)](s, \pi_h^k(s)). \end{aligned}$$

We bound the terms involving $V_{h+1}^{\pi^k}$ using our transition bounds. Using optimism (Lemma 6) and bounded loss Assumption 3, we have $0 \leq V_{h+1}^{\pi^k} - \hat{V}_{h+1}^k \leq H$. Thus, the good event (Eq. (8)) gives:

$$\hat{P}_h^{k,b}(V_{h+1}^{\pi^k} - \hat{V}_{h+1}^k) \leq \left(1 + \frac{1}{3H}\right) P_h(V_{h+1}^{\pi^k} - \hat{V}_{h+1}^k) + 2.2H^2S\kappa\beta_h^k.$$

Similarly, $0 \leq V_{h+1}^{\pi^k} - V_{h+1}^\star \leq H$, so applying Eq. (9) and the optimism property yields:

$$\begin{aligned} (P_h - \hat{P}_h^{k,b})(V_{h+1}^{\pi^k} - V_{h+1}^\star) &\leq \frac{1}{3H} P_h(V_{h+1}^{\pi^k} - V_{h+1}^\star) + 2.2H^2S\kappa\beta_h^k \\ &\leq \frac{1}{3H} P_h(V_{h+1}^{\pi^k} - \hat{V}_{h+1}^k) + 2.2H^2S\kappa\beta_h^k. \end{aligned}$$

Plugging the last two inequalities into Eq. (11) gives

$$\begin{aligned} V_h^{\pi^k}(s) - \hat{V}_h^{k,b}(s) \leq\ & [(\ell_h - \hat{\ell}_h^{k,b}) + (P_h - \hat{P}_h^{k,b})V_{h+1}^\star](s, \pi_h^k(s)) \\ &+ [4.4H^2S\kappa\beta_h^k + (1 + 2/(3H))P_h(V_{h+1}^{\pi^k} - \hat{V}_{h+1}^k)](s, \pi_h^k(s)). \end{aligned}$$

Now, notice that $\hat{V}_h^k = q_\alpha(\hat{V}_h^{k,b}, b \in [B])$ (see Algorithm 1) and that the quantile (Eq. (1)) is equivariant to shifts, i.e., $q_\alpha(\mu_b + c, b \in [B]) = q_\alpha(\mu_b, b \in [B]) + c$. Applying this with the good event's Eq. (7), and noticing that $\kappa \geq \log 40$ we get the step-wise recursion:

$$\begin{aligned} V_h^{\pi^k}(s) - \hat{V}_h^k(s) \leq\ & 1.7\sqrt{\beta_h^k(s, \pi_h^k(s)) \operatorname*{Var}_{s,\pi_h^k(s),h}(L_h + V_{h+1}^\star(s_{h+1}))} + 7H^2S\kappa\beta_h^k(s, \pi^k(s)) \\ &+ (1 + 2/(3H))P_h(V_{h+1}^{\pi^k} - \hat{V}_{h+1}^k)(s, \pi_h^k(s)). \end{aligned} \qquad (12)$$

Recalling that $\mathrm{Var}_{s,\pi_h^k(s),h}(L_h + V_{h+1}^\star(s_{h+1})) \le \mathbb{Q}_h^\star$ (see Eq. (4)), and unrolling the recursion, we get that for any $\eta > 0$

$$V_1^{\pi^k}(s_1) - \hat{V}_1^k(s_1) \le \mathbb{E}_{P,\pi^k} \sum_{h\in[H]} (1+1/H)^{h-1} \left[1.7\sqrt{\mathbb{Q}_h^\star \beta_h^k(s_h,a_h)} + 7H^2 S\kappa \beta_h^k(s_h,a_h)\right]$$
$$\le \eta^{-1}\mathbb{Q}^\star + (5.5\eta + 20H^2 S\kappa)\mathbb{E}_{P,\pi^k} \sum_{h\in[H]} \beta_h^k(s_h,a_h),$$

where the second inequality used $(1+1/H)^h \le e$ for all $h \in [H]$, and the AM-GM inequality ($2\sqrt{ab} \le a+b$). This completes the first part of the lemma.

Next, by optimism (Lemma 6) and the definition of $V^\star$, we have $\hat{V}_{h+1}^k \le V_{h+1}^\star \le V^{\pi^k}$. Thus, plugging Lemma 10 into Eq. (12), we get that for any $\eta > 0$

$$V_h^{\pi^k}(s) - \hat{V}_h^k(s) \le \eta^{-1} \operatorname*{Var}_{s,\pi_h^k(s),h}(L_h + V_{h+1}^{\pi^k}(s_{h+1})) + ((2/e)\eta + 8H^2 S\kappa)\beta_h^k(s,\pi^k(s))$$
$$+ (1+1/H)P_h(V_{h+1}^{\pi^k} - \hat{V}_{h+1}^k)(s,\pi^k(s)).$$

Choosing $\eta \ge 5eH^2$ and similarly unrolling the recursion, we have

$$\begin{aligned}
&V_1^{\pi^k}(s_1) - \hat{V}_1^k(s_1) \\
&\le \mathbb{E}_{P,\pi^k} \sum_{h\in[H]} (1+1/H)^{h-1} \left[\eta^{-1} \operatorname*{Var}_{s,\pi_h^k(s),h}(L_h + V_{h+1}^{\pi^k}(s_{h+1})) + ((2/e)\eta + 8H^2 S\kappa)\beta_h^k(s,\pi^k(s))\right] \\
&\le e\eta^{-1} H V_1^{\pi^k}(s_1) + (2\eta + 22H^2 S\kappa)\mathbb{E}_{P,\pi^k} \sum_{h\in[H]} \beta_h^k(s_h,a_h) \qquad \text{(Lemma 12)} \\
&\le (V_1^{\pi^k}(s_1) - V^\star)/5 + e\eta^{-1} H V^\star + (2\eta + 22H^2 S\kappa)\mathbb{E}_{P,\pi^k} \sum_{h\in[H]} \beta_h^k(s_h,a_h),
\end{aligned}$$

where Lemma 12 is an application of the Law of Total Variance (Azar et al., 2017) together with the bounded total loss assumption. Using that $\hat{V}_1^k(s_1) \le V^\star$ (Lemma 6) and rearranging, we have

$$V_1^{\pi^k}(s_1) - \hat{V}_1^k(s_1) \le 3.4\eta^{-1} H V^\star + (2.5\eta + 27.5H^2 S\kappa)\mathbb{E}_{P,\pi^k} \sum_{h\in[H]} \beta_h^k(s_h,a_h).$$

Replacing η with $3.4\eta + 5eH^2$ and using $\kappa \ge \log 40 \approx 3.68$ concludes the proof. ■

7 Conclusion, Limitations and Future Work

In this work, we presented *VIBE*, establishing the first theoretical guarantee that an ensemble of value function estimators over partitioned data provably achieves minimax-optimal, variance-dependent regret without explicit exploration bonuses or posterior sampling. Our analysis provides a formal foundation for the empirical success of ensemble exploration heuristics in deep RL (Osband et al., 2016a; Lee et al., 2021), demonstrating that optimism in the face of uncertainty emerges naturally through quantile aggregation. A notable limitation of our finite-sample guarantees lies in the conservative lower-order constants dictated by Feige's inequality (Feige, 2004). Despite this conservative scaling requirement, *VIBE* performs well out of the box in our empirical evaluations (Section D), while showing the potential to surpass leading baselines when scaled according to Feige's tighter conjectured constant ($1/e$ instead of $1/13$).

Crucially, while our formal guarantees focus on the tabular regime, the algorithmic recipe of *VIBE* is immediately practical for large-scale challenges. Unlike bonus-based methods that require intractable density estimation, pseudo-counts, or representation-space covariance tracking, quantile-based exploration can be directly instantiated in deep RL simply by training an ensemble of neural networks on mini-batches and selecting actions via quantile aggregation. Future directions on the theoretical side include establishing formal regret bounds under function approximation (addressing Bellman error propagation under quantile operators) and rigorously characterizing the impact of stochastic optimization noise arising from temporal-difference learning and SGD over replay buffers.

AI USE STATEMENT

AI tools were used to assist with drafting and editing the manuscript, as well as generating code for the empirical evaluations. The authors have reviewed and verified all text and code, and take full responsibility for the content.

ETHICS STATEMENT

This paper presents foundational theoretical research in reinforcement learning and multi-armed bandits. As a purely theoretical study without human subjects or sensitive datasets, it has no immediate negative ethical or societal implications.

REPRODUCIBILITY STATEMENT

All theoretical claims and regret bounds are accompanied by complete, self-contained mathematical proofs in the appendix. The proposed algorithms are detailed with full pseudocode, exact hyperparameters, and implementation specifications. Furthermore, complete experimental details, environment specifications, and baseline configurations are provided in Section D, and the full source code to reproduce all empirical results is included in the supplementary material.

REFERENCES

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in neural information processing systems*, 30, 2017.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pp. 263–272. PMLR, 2017.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26. JMLR Workshop and Conference Proceedings, 2011.

Asaf Cassel, Orin Levy, and Yishay Mansour. Batch ensemble for variance dependent regret in stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15678–15685, 2025.

Richard Y Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. Ucb exploration via q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith W. Ross. Randomized ensembled double q-learning: Learning fast without a model. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1507–1516. PMLR, 09–15 Jun 2019.

Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1–12. Curran Associates, Inc., 2021.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato (eds.), *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 578–598. PMLR, 16–19 Mar 2021.

Uriel Feige. On sums of independent random variables with unbounded variance, and estimating the average degree in a graph. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pp. 594–603, 2004.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. ISSN 00911798, 2168894X.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. PMLR, 2018.

Botao Hao, Yasin Abbasi Yadkori, Zheng Wen, and Guang Cheng. Bootstrapping upper confidence bound. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Emilie Kaufmann, Olivier Cappe, and Aurelien Garivier. On bayesian upper confidence bounds for bandit problems. In Neil D. Lawrence and Mark Girolami (eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 592–600, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

Branislav Kveton, Csaba Szepesvári, Sharan Vaswani, Zheng Wen, Tor Lattimore, and Mohammad Ghavamzadeh. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3601–3610. PMLR, 2019.

Cassidy Laidlaw, Stuart J Russell, and Anca Dragan. Bridging rl theory and practice with the effective horizon. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 58953–59007. Curran Associates, Inc., 2023. doi: 10.52202/075280-2573.

Harin Lee and Min-hwan Oh. Improved regret of linear ensemble sampling. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 92803–92831. Curran Associates, Inc., 2024. doi: 10.52202/079017-2947.

Harin Lee and Min-hwan Oh. Minimax optimal reinforcement learning with quasi-optimism. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 13184–13231, 2025.

Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *International conference on machine learning*, pp. 6131–6141. PMLR, 2021.

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. *Advances in neural information processing systems*, 30, 2017.

Nadav Merlis. Reinforcement learning with lookahead information. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 64523–64581. Curran Associates, Inc., 2024.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016a.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pp. 2377–2386. PMLR, 2016b.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International conference on machine learning*, pp. 5062–5071. PMLR, 2019.

Oren Peer, Chen Tessler, Nadav Merlis, and Ron Meir. Ensemble bootstrapping for q-learning. In *International Conference on Machine Learning*, pp. 8454–8463. PMLR, 2021.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pp. 8210–8219. PMLR, 2020.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *Advances in neural information processing systems*, 32, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Munos, Alexey Naumov, Mark Rowland, Michal Valko, and Pierre Ménard. Optimistic posterior sampling for reinforcement learning with few samples and tight guarantees. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10737–10751. Curran Associates, Inc., 2022a.

Daniil Tiapkin, Denis Belomestny, Eric Moulines, Alexey Naumov, Sergey Samsonov, Yunhao Tang, Michal Valko, and Pierre Menard. From Dirichlet to rubin: Optimistic exploration in RL without bonuses. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21380–21431. PMLR, 17–23 Jul 2022b.

Stefano Viel, Luca Viano, and Volkan Cevher. IL-SOAR : Imitation learning with soft optimistic actor cRitic. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61444–61479. PMLR, 13–19 Jul 2025.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7304–7312. PMLR, 09–15 Jun 2019.

Runlong Zhou, Zhang Zihan, and Simon Shaolei Du. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42878–42914. PMLR, 23–29 Jul 2023.

A Additional Proofs

A.1 Proof of Lemma 1 (QoM)

For the first part, notice that the event holds when at least $\lceil \alpha B \rceil$ batches yield an estimate less than or equal to μ. By Corollary 2, each batch satisfies this with probability at least $1/13$. Because the subsets $\mathcal{D}^b$ are fixed and disjoint, they are independent. Letting $\mathrm{Bin}(n, p)$ be the binomial distribution with n trails and success probability p, we conclude that

$$\begin{aligned}\Pr(\hat{\mu}_\alpha \le \mu) \ge \Pr(\mathrm{Bin}(B, 1/13) \ge \lceil B/65 \rceil) &\ge \Pr(\mathrm{Bin}(B, 1/13) > B/65) \\ &= 1 - \Pr(\mathrm{Bin}(B, 1/13) \le B/65) \\ &\ge 1 - \exp(-B \cdot \mathrm{D}_{\mathrm{KL}}(1/65 || 1/13)) \\ &\ge 1 - \exp(-B/26) \\ &\ge 1 - \delta,\end{aligned}$$

where the fourth transition used a standard Chernoff bound for Binomial random variables, and D_{KL} is the Kullback–Leibler divergence.

The second part of the proof follows similar logic. Denoting the mean estimate of batch b in Eq. (2) as $\hat{\mu}^b$, we use Lemma 14 (Freedman's inequality) to get

$$\Pr\left(\sum_{X \in \mathcal{D}^b} X \ge |\mathcal{D}^b|\mu - 1.7\sqrt{|\mathcal{D}^b|\sigma^2} - 8R\right) \ge 0.99953.$$

Rearranging the inner term, dividing by $|\mathcal{D}^b| + 1 \ge B^{-1}\max\{1, n\}$ we get

$$\Pr\left(\hat{\mu}^b \ge \mu - 1.7\sqrt{\frac{\sigma^2 B}{\max\{1, n\}}} - \frac{9RB}{\max\{1, n\}}\right) \ge 0.99953.$$

For the event to hold, at least $B + 1 - \lceil B/65 \rceil$ batches need to satisfy the above. As before, this occurs with probability at least

$$\begin{aligned}\Pr(\mathrm{Bin}(B, 0.99953) \ge B + 1 - \lceil B/65 \rceil) &\ge \Pr(\mathrm{Bin}(B, 0.99953) > 64B/65) \\ &= 1 - \Pr(\mathrm{Bin}(B, 0.99953) \le 64B/65) \\ &\ge 1 - \exp(-B \cdot \mathrm{D}_{\mathrm{KL}}(64/65 || 0.99953)) \\ &\ge 1 - \exp(-B/26) \\ &\ge 1 - \delta. \qquad (\blacksquare)\end{aligned}$$

A.2 Proof of Lemma 7 (Good Event)

Because of the round-robin schedule (Eq. (3)), the samples from each state-action pair are (nearly) uniformly distributed between the datasets. Thus, we can relate the size of each dataset with the global counts as $|D_h^{k,b}(s,a)| \ge \lfloor N_h^k(s,a)/B \rfloor$. Therefore, for all $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], k \in [K]$, and $b \in [B]$ we have

$$|D_h^{k,b}(s,a)| + 1 \ge B^{-1}\max\{1, N_h^k(s,a)\}. \tag{13}$$

Now, we show that each of Eqs. (6) to (10) hold with probability at least $1 - \delta/5$; thus, taking a union bound concludes the proof of Lemma 7.

(Eq. (6)): Follows from Lemma 5.

(Eq. (7)): As shown in Lemma 5, $\hat{\mathcal{T}}_h^k V_{h+1}^\star(s,a)$ satisfies the basic conditions of Lemma 1. To satisfy the additional conditions in the second claim of Lemma 1, we use the bounded total loss assumption to bound each sample by H, and Eq. (13) to bound the size of the datasets. Thus, the proof is concluded via a union bound on $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$, and the (at most K) dataset configurations (see full explanation in the proof of Lemma 5).

(Eqs. (8) and (9)): Fix $s', s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], b \in [B], |\mathcal{D}_h^{k,b}(s,a)| \in [K/B]$. For the first inequality, we apply Lemma 14 with $X_t = \mathbb{1}\{s_+ = s'\} - p_h(s' \mid s,a)$, for $s_+ \in \mathcal{D}_h^{k,b}(s,a)$ and $\lambda = 1/3(e-2)H$. Notice that $R = 1$ and

$$\mathbb{E}[X_t^2 \mid X_1, \dots, X_{t-1}] = P_h(s' \mid s,a)(1 - P_h(s' \mid s,a)) \le P_h(s' \mid s,a),$$

where the last inequality holds since $P_h \in [0,1]$. Thus, by Lemma 14, with probability at least $1-\delta$

$$\sum_{s_+ \in \mathcal{D}_h^{k,b}(s,a)} \mathbb{1}\{s_+ = s'\} - P_h(s' \mid s,a) \le \frac{1}{3H}|\mathcal{D}_h^{k,b}(s,a)|p_h(s' \mid s,a) + 3(e-2)H \log \frac{1}{\delta}.$$

Recalling the definition of $\hat{P}$ in Algorithm 1, rearranging the inequality and dividing by $|\mathcal{D}_h^{k,b}(s,a)| + 1$ gives

$$\hat{P}_h^{k,b}(s' \mid s,a) - \left(1 + \frac{1}{3H}\right) P_h(s' \mid s,a) \le \frac{3(e-2)H \log \delta^{-1}}{|\mathcal{D}_h^{k,b}(s,a)| + 1} \le \frac{3(e-2)HB \log \delta^{-1}}{\max\{1, N_h^k(s,a)\}}, \quad (14)$$

where the last inequality used Eq. (13). For the second inequality, we apply Lemma 14 with $X_t = P_h(s' \mid s,a) - \mathbb{1}\{s_+ = s'\}$ to get that with probability at least $1-\delta$

$$\sum_{s_+ \in \mathcal{D}_h^{k,b}(s,a)} P_h(s' \mid s,a) - \mathbb{1}\{s_+ = s'\} \le \frac{1}{3H}|\mathcal{D}_h^{k,b}(s,a)|p_h(s' \mid s,a) + 3(e-2)H \log \frac{1}{\delta}.$$

Adding $P_h(s' \mid s,a)$ to both sides, rearranging and dividing by $|\mathcal{D}_h^{k,b}(s,a)| + 1$ gives

$$\left(1 - \frac{1}{3H}\right) P_h(s' \mid s,a) - \hat{P}_h^{k,b}(s' \mid s,a) \le \frac{3(e-2)H \log e\delta^{-1}}{|\mathcal{D}_h^{k,b}(s,a)| + 1} \le \frac{3(e-2)HB \log e\delta^{-1}}{\max\{1, N_h^k(s,a)\}}, \quad (15)$$

where the last inequality used Eq. (13). Setting δ in Eqs. (14) and (15) appropriately and performing a union bound over $s', s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], b \in [B], |\mathcal{D}_h^{k,b}(s,a)| \in [K/B]$ concludes the proof.

(Eq. (10)): First, applying Lemma 15 with $X_t = \sum_{h \in [H]} \beta_h^k(s_h, a_h) \le HB$ gives that with probability at least $1 - \delta/5$

$$\sum_{k \in [K]} \mathbb{E}_{\pi^k, P} \sum_{h \in [H]} \beta_h^k(s_h, a_h) \le 2 \sum_{k \in [K]} \sum_{h \in [H]} \beta_h^k(s_h^k, a_h^k) + 4HB \log 5\delta^{-1}.$$

The proof is concluded by applying Lemma 9, a standard algebraic argument on the harmonic sum of the state-action counts, together with the observation that the regret is non-zero only if $A \ge 2$. ■

A.3 Existing Results

A.3.1 Summing the counts

Lemma 9. *Recall the state-action counts $N_h^k(s,a) = \sum_{k' \in [k-1]} \mathbb{1}\Big\{(s,a) = (s_h^{k'}, a_h^{k'})\Big\}$. We have that*

$$\sum_{k \in [K]} \sum_{h \in [H]} \frac{1}{\max\{1, N_h^k(s_h^k, a_h^k)\}} \le SAH \log 8K.$$

Proof (for sake of completeness). Recall that $N_h^k(s,a) = \sum_{k' \in [k-1]} \mathbb{1}\Big\{(s_h^{k'}, a_h^{k'}) = (s,a)\Big\}$, i.e., the counts are incremented every time a state-action pair is visited. Thus, we have

$$\begin{aligned}
\sum_{k\in[K]} \sum_{h\in[H]} \frac{1}{\max\{1, N_h^k(s_h^k, a_h^k)\}} &= \sum_{h\in[H]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{k\in[K]} \frac{\mathbb{1}\big\{(s_h^k, a_h^k) = (s,a)\big\}}{\max\{1, N_h^k(s_h^k, a_h^k)\}} \\
&= \sum_{h\in[H]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{i=0}^{N_h^K(s,a)} \frac{1}{\max\{1,i\}} \\
&\le \sum_{h\in[H]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{i=0}^{K} \frac{1}{\max\{1,i\}} \\
&\le \sum_{h\in[H]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \log 8K \\
&= SAH \log 8K,
\end{aligned}$$

where the first inequality used that $N_h^k(s,a) \le K$ for all $k \in [K]$. ■

A.3.2 Variance Manipulation

The following result is a variant of Lemma 21 in Merlis (2024).

Lemma 10. *Let $V, \overline{V}, \underline{V} \in [0,H]^{\mathcal{S}}$ be value functions such that $\underline{V}(s) \le V(s) \le \overline{V}(s)$ for all $s \in \mathcal{S}$. Then for any $h \in [H], s \in \mathcal{S}, a \in \mathcal{A}, \beta, \eta > 0$*

$$\sqrt{\beta \operatorname*{Var}_{s,a,h}(L_h + V(s_{h+1}))} \le \eta^{-1} \operatorname*{Var}_{s,a,h}(L_h + \overline{V}(s_{h+1})) + \frac{1}{3H} P_h(\overline{V} - \underline{V})(s_h, a_h) + \frac{(3H^2 + \eta)\beta}{4}.$$

Proof. Recall that the standard deviation satisfies a triangle inequality (see Zanette & Brunskill (2019) Eqs. 48-51), i.e., for $\sigma(X) = \sqrt{\mathrm{Var}(X)}$ we have $\sigma(X+Y) \le \sigma(X) + \sigma(Y)$. Thus, we have

$$\sqrt{\beta \operatorname*{Var}_{s,a,h}(L_h + V(s_{h+1}))} \le \sqrt{\beta \operatorname*{Var}_{s,a,h}(L_h + \overline{V}(s_{h+1}))} + \sqrt{\beta \operatorname*{Var}_{s,a,h}((V - \overline{V})(s_{h+1}))}.$$

Now, using our assumption on the value functions, we have

$$\begin{aligned}
\sqrt{\beta \operatorname*{Var}_{s,a,h}((V - \overline{V})(s_{h+1}))} \le \sqrt{\beta H P_h(\overline{V} - V)(s,a)} &\le \sqrt{\beta H P_h(\overline{V} - \underline{V})(s,a)} \\
&\le (1/3H) P_h(\overline{V} - \underline{V})(s,a) + (3/4) H^2 \beta,
\end{aligned}$$

where the last step also used the AM-GM inequality ($\sqrt{ab} \le (a^2 + b^2)/2$). Plugging this into the previous inequality and using AM-GM on the first term concludes the proof. ■

Next, we state the well-known Law of Total Variance (LTV), originally due to Azar et al. (2017).

Lemma 11 (Zanette & Brunskill (2019), Lemma 15). *For any policy π, it holds that*

$$\mathbb{E}_{P,\pi}\left[\sum_{h\in[H]} \operatorname*{Var}_{s_h,a_h,h}(V^\pi(s_{h+1}))\right] = \mathbb{E}_{P,\pi}\left[\sum_{h\in[H]} \ell_h(s_h,a_h) - V_1^\pi(s_1)\right]^2$$

Lemma 12 (Variance Bound). *For any policy $\pi : \mathcal{S} \to \mathcal{A}$ we have*

$$\mathbb{E}_{P,\pi} \sum_{h\in[H]} \operatorname*{Var}_{s_h,a_h,h}(L_h + V_{h+1}^\pi(s_{h+1})) = \mathbb{E}_{P,\pi,L}\left[\sum_{h\in[H]} L_h - V_1^\pi(s_1)\right]^2 \le H V_1^\pi(s_1).$$

Proof. Recall the bounded total loss assumption, $0 \leq \sum_{h\in[H]} L_h \leq H$. Thus, we have

$$\mathbb{E}_{P,\pi,L}\left[\sum_{h\in[H]} L_h - V_1^\pi(s_1)\right]^2 \leq \mathbb{E}_{P,\pi,L}\left[\sum_{h\in[H]} L_h\right]^2 \leq H\mathbb{E}_{P,\pi,L}\left[\sum_{h\in[H]} L_h\right] = HV_1^\pi(s_1).$$

Next, notice that s_{h+1} and L_h are conditionally independent on s_h, a_h, thus

$$\begin{aligned}\mathbb{E}_{P,\pi}\sum_{h\in[H]} \underset{s_h,a_h,h}{\mathrm{Var}}(L_h+V_{h+1}^\pi(s_{h+1})) &= \mathbb{E}_{P,\pi}\sum_{h\in[H]} \underset{s_h,a_h,h}{\mathrm{Var}}(L_h) + \underset{s_h,a_h,h}{\mathrm{Var}}(V_{h+1}^\pi(s_{h+1})) \\ &= \mathbb{E}_{P,\pi}\left(\sum_{h\in[H]} \underset{s_h,a_h,h}{\mathrm{Var}}(L_h)\right) + \left(\sum_{h\in[H]} \ell_h(s_h,a_h) - V_1^\pi(s_1)\right)^2,\end{aligned}$$

where the second equality used Lemma 11 (LTV). Now, let $\iota = ((s_h,a_h)_{h\in[H]}$ be a trajectory. Because π does not depend on the losses $L_h, h \in [H]$ we have $\mathrm{Var}_{s_h,a_h,h}(L_h) = \mathbb{E}_L[(L_h - \ell_h(s_h,a_h))^2 \mid \iota]$. Furthermore, conditioned on ι, the losses are independent, thus $\mathbb{E}_{P,\pi}\sum_{h\in[H]}\mathrm{Var}_{s_h,a_h,h}(L_h) = \mathbb{E}_{P,\pi,L}\left(\sum_{h\in[H]} L_h - \ell_h(s_h,a_h)\right)^2$. Plugging into the above and showing that the two terms are uncorrelated concludes the proof. To see the latter, notice that

$$\begin{aligned}&\mathbb{E}_{P,\pi,L}\left(\sum_{h\in[H]} L_h - \ell_h(s_h,a_h)\right)\left(\sum_{h\in[H]} \ell_h(s_h,a_h) - V_1^\pi(s_1)\right) \\ &\quad= \mathbb{E}_{\iota\sim P,\pi}\mathbb{E}_L\left[\left(\sum_{h\in[H]} L_h - \ell_h(s_h,a_h)\right)\middle|\iota\right]\left(\sum_{h\in[H]} \ell_h(s_h,a_h) - V_1^\pi(s_1)\right) = 0. \quad \blacksquare\end{aligned}$$

A.3.3 Statistical Inequalities

Lemma 13 (Feige (2004), Theorem 1). *Let $X_1,\ldots,X_n$ be arbitrary nonnegative independent random variables, with expectations $\mu_1,\ldots,\mu_n$ respectively, where $\mu_i \leq 1$ for every i. Let $X = \sum_{i=1}^n X_i$, and let μ denote the expectation of X (hence, $\mu = \sum_{i=1}^n \mu_i$). Then for every $\delta > 0$*

$$\Pr(X < \mu + \delta) \geq \min\{\delta/(1+\delta), 1/13\}.$$

The following is a Freedman-type inequality Freedman (1975) derived by Beygelzimer et al. (2011).

Lemma 14 (Beygelzimer et al. (2011), Theorem 1). *Let $X_1, X_2, \ldots, X_T$ be a sequence of real-valued random variables. Assume for all $t \in [T]$, $X_t \leq R$ and $\mathbb{E}[X_t \mid X_1,\ldots,X_{t-1}] = 0$. Then for any $\delta \in (0,1)$ and $\lambda \in [0, 1/R]$, with probability at least $1-\delta$*

$$\sum_{t\in[T]} X_t \leq (e-2)\lambda \sum_{t\in[T]} \mathbb{E}[X_t^2 \mid X_1,\ldots,X_{t-1}] + \lambda^{-1}\log\frac{1}{\delta}.$$

If $\mathbb{E}[X_t^2 \mid X_1,\ldots,X_{t-1}] = \sigma^2$ then, setting $\lambda = \min\{R^{-1}, \sqrt{((e-2)\sigma^2 T)^{-1}\log\delta^{-1}}\}$, gives

$$\sum_{t\in[T]} X_t \leq 2\sqrt{(e-2)\sigma^2 T} + R\log\frac{1}{\delta}.$$

Next, we state the following Bernstein-type tail bound (e.g.,Rosenberg et al. (2020), Lemma D.4).

Lemma 15. *Let $\{X_t\}_{t\geq 1}$ be a sequence of random variables with expectation adapted to a filtration $\mathcal{F}_t$. Suppose that $0 \leq X_t \leq 1$ almost surely. Then with probability at least $1-\delta$*

$$\sum_{t=1}^T \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \leq 2\sum_{t=1}^T X_t + 4\log\frac{1}{\delta}.$$

B Extension to Heavy-Tailed Losses

In this section we discuss how Theorem 4 extends to losses with heavy tails. Specifically, we can replace Assumption 3 with the weaker requirement that the random losses L_h are non-negative, satisfy $\mathbb{E}\left[\sum_{h\in[H]} L_h\right] \le H$ for all policies, and each $L_h(s,a)$ has finite (but unknown) variance.

The key observation enabling this extension is that Lemma 1 applies to non-negative random variables with finite variance, without requiring boundedness (as discussed following its statement). Consequently, the QoM estimator remains optimistic and its bias increases by at most a constant multiplicative factor in this setting. Given this, the good event used in the proof of Theorem 4 holds essentially unchanged, and the remainder of the regret decomposition follows straightforwardly, yielding the same regret bound up to constants.

We note, however, that the $\min\{\mathbb{Q}^\star, HV^\star\}$ term appearing in Theorem 4 reduces to $\mathbb{Q}^\star$ alone in the heavy-tailed setting. The $HV^\star$ term arises from bounding the variance of the value function under the optimal policy, which relies on the losses being almost surely bounded. Since no such bound is assumed in the heavy-tailed setting, this refinement does not apply, and we cannot expect the $HV^\star$ term to appear in the regret bound.

C Multi-Armed Bandits

Problem setup. A stochastic multi-armed bandit is an MDP with a single state and horizon one, i.e., $S = H = 1$ (see Section 3). Thus, we simplify the notation as follows. In each episode, $k \in [K]$ the agent chooses an arm $a^k \in \mathcal{A}$ and observes a stochastic loss $L^k \in [0,1]$ sampled independently from a distribution $\mathcal{L}(a^k)$ whose mean and variance are denoted $\ell(a^k), \sigma^2(a^k)$. Denoting the optimal action and value as $a^\star \in \arg\min_{a\in\mathcal{A}} \ell(a)$ and $\ell^\star = \mu(a^\star)$, the (pseudo) regret can be rewritten as

$$\mathrm{regret}_K = \sum_{k\in[K]} [\ell(a^k) - \ell^\star] = \sum_{k\in[K]} \Delta(a^k),$$

where $\Delta(a) = \ell(a) - \ell^\star$ is known as the sub-optimality gap of arm $a \in \mathcal{A}$.

C.1 Algorithm and Main Results

We present Algorithm 2, which is identical to Algorithm 1 but with a simplified notation.

Algorithm 2 Quantile of Mean for MAB

1: **input**: Number of batches B, quantile level α.
2: **initialize**: $N^1(a) = 0, \mathcal{D}^{1,b}(a) = \emptyset$ for all $a \in A$.
3: **for** episode $k = 1, 2, \ldots, K$ **do**

$$\hat{\ell}^{k,b}(a) = \sum_{L\in\mathcal{D}^{k,b}(a)} \frac{L}{|\mathcal{D}^{k,b}(a)| + 1}$$
$$\hat{\ell}^{k}(a) \;\; = q_\alpha\Big(\hat{\ell}^{k,b}(a), b \in [B]\Big).$$

4: Play $a^k \in \arg\min_{a\in\mathcal{A}} \hat{\ell}^k(a)$ and observe loss L^k.
5: Update global counts: $N^{k+1}(a) = N^k(a) + \mathbb{1}\{a = a^k\}$
6: Calculate batch index: $b^k = 1 + (N^k(a^k) \mod B)$ and update datasets:

$$\mathcal{D}^{k+1,b}(a) = \mathcal{D}^{k,b}(a) \bigcup \begin{cases} L^k, & \text{if } (a,b) = (a^k, b^k) \\ \emptyset, & \text{otherwise.} \end{cases}$$

The following is our main result for multi-armed bandits (proof in Section C.2).

Theorem 16. *Suppose we run Algorithm 2 with number of batches $B = 26\log(2KA\delta^{-1})$, and quantile level $\alpha = 1/65$. With probability at least $1-\delta$, the following hold simultaneously*

$$\mathrm{regret}_K \le \sqrt{76K\sum_{a\neq a^\star}\sigma^2(a)\log(6KA\delta^{-1})} + 469A\log(2KA\delta^{-1})$$
$$\mathrm{regret}_K \le \sum_{a\neq a^\star}\left(\frac{76\sigma^2(a)}{\Delta(a)} + 469\right)\log(2KA\delta^{-1}).$$

We note that Feige (2004) conjecture that their result, stated here as Lemma 13, holds with a constant of $1/e$. If true, this would change the numerical constant in the batch size to 5.5, and a better choice of α would be $1/10$. Overall, this would decrease the constants in our bounds. Finally, we note that this result does not require the losses to be bounded in $[0,1]$. In fact, it holds even for nonnegative losses with finite variance, albeit with significantly larger numerical constants. To see this, replace the use of Freedman's inequality in the proof of Lemma 1 with Chebyshev's inequality.

C.2 Proof of Theorem 16

Recall that the pseudo-regret may be written as

$$\mathrm{regret}_K = \sum_{a\neq a^\star} N^{K+1}(a)\Delta(a), \tag{16}$$

where $N^k(a)$ is defined in Algorithm 2. Thus, our goal is to bound $N^{K+1}(a)$ for each sub-optimal arm. We begin with a standard "good event" over which the regret is bounded deterministically. Suppose that for all $k\in[K]$ and $a\neq a^\star$ we have

$$\hat{\ell}^k(a^\star) \le \ell^\star \tag{17}$$
$$\hat{\ell}^k(a) \ge \ell(a) - \frac{9B}{N^k(a)} - 1.7\sqrt{\frac{\sigma^2(a)B}{N^k(a)}}. \tag{18}$$

Taking a union bound over Lemma 1 with $\delta/2KA$, the above holds with probability at least $1-\delta$. Now, suppose that arm a with $\Delta(a)>0$ was played at episode k. Then, by the decision rule in Algorithm 2, we have $\hat{\ell}^k(a^\star)\ge\hat{\ell}^k(a)$, and thus

$$\begin{aligned}\ell^\star &\ge \hat{\ell}^k(a^\star) && \text{(Eq. (17))}\\ &\ge \hat{\ell}^k(a)\\ &\ge \ell(a) - \frac{9B}{N^k(a)} - 1.7\sqrt{\frac{\sigma^2(a)B}{N^k(a)}}. && \text{(Eq. (18))}\end{aligned}$$

Solving this quadratic inequality for $N^k(a)$, we have

$$N^k(a) \le B\left[\frac{2.9\sigma^2(a)}{\Delta^2(a)} + \frac{18}{\Delta(a)}\right],$$

Now, let k_a be the last episode when arm a was chosen. Then we have

$$N^{K+1}(a) = N^{k_a}(a) + 1 \le \left(\frac{76\sigma^2(a)}{\Delta^2(a)} + \frac{469}{\Delta(a)}\right)\log(2KA\delta^{-1}),$$

where the inequality also used our choice of $B = 26\log(2KA\delta^{-1})$. Plugging this into Eq. (16) concludes the instance-dependent regret bounds.

Next, to obtain the instance-independent bounds, notice that

$$\begin{aligned}N^{K+1}(a)\Delta(a) &\le \min\left\{N^{K+1}(a)\Delta(a), \frac{76\sigma^2(a)}{\Delta(a)}\log(2KA\delta^{-1})\right\} + 469\log(2KA\delta^{-1})\\ &\le \sqrt{76N^{K+1}(a)\sigma^2(a)\log(2KA\delta^{-1})} + 469\log(2KA\delta^{-1}),\end{aligned}$$

where the second inequality used $\min\{c_1,c_2\}\le\sqrt{c_1c_2}$. Plugging this into Eq. (16) and applying the Cauchy-Schwarz inequality concludes the proof. ■

D Experimental Setup and Empirical Results

While this work primarily focuses on establishing theoretical guarantees for bonus-free ensemble exploration in MDPs, we provide empirical proof-of-concept evaluations to investigate how *VIBE* (Algorithm 1) compares with standard exploration methods in practice. To ensure a consistent comparison, we adopt the experimental harness and baseline suite of Lee & Oh (2025). Our implementation builds upon their published codebase, extending it where key experimental details were omitted or unclear—specifically in formalizing action masking for state-dependent action spaces and evaluating algorithms strictly under their theoretical closed-form parameters. All code and environment configurations are provided in the supplementary material. Across all domains, algorithms are evaluated under the non-uniform reward regime with failure probability $\delta = 0.05$, with results averaged over 10 independent runs.

Benchmark Environments.

- **RiverSwim:** A classic chain of S states with $A = 2$ actions (Lee & Oh, 2025). Swimming left moves downstream deterministically. Swimming right attempts to move upstream against a stochastic current: from intermediate states, the agent has a 35% probability of advancing right, a 60% probability of remaining in place, and a 5% probability of slipping back left (at the boundaries, the agent advances or stays with probability 0.6, slipping with probability 0.4). A small reward $r = 0.005$ is placed at s_1 and a goal reward $r = 1.0$ at s_S. We evaluate four scaling configurations: $(S, H) \in \{(10, 40), (20, 80), (30, 120), (40, 160)\}$ over $K = 100{,}000$ episodes.
- **Gap MDP:** Introduced by Dann et al. (2021) to evaluate whether algorithms can leverage large value gaps along the optimal trajectory ($\Delta = 0.5$) to achieve fast, instance-dependent (logarithmic) regret, rather than suffering worst-case regret exploring n suboptimal branches due to their small local gap ($\varepsilon = 0.02$). In this 3-layer MDP ($H = 3$), root action a_1 reaches an optimal state yielding $r \sim \mathrm{Bernoulli}(0.5)$, while n distracting actions $a_2, \ldots, a_{n+1}$ lead to parallel suboptimal states yielding $r \sim \mathrm{Bernoulli}(0.02)$. We evaluate $n = 16$ ($S = 22, A = 17$) and $n = 32$ ($S = 38, A = 33$) over $K = 500{,}000$ episodes.
- **BRIDGE Environments:** Complex tabularized representations from Laidlaw et al. (2023): `Freeway-10` ($S = 198, A = 3, H = 10$) and `KeyCorridor-S3R1` ($S = 169, A = 6, H = 20$), evaluated over $K = 500{,}000$ episodes.

Tested VIBE Configurations. We evaluate two parameter choices for *VIBE*:

- **Theoretical VIBE (Vanilla):** Uses the conservative theoretical parameters from Theorem 4: $\alpha = 1/65 \approx 0.0154$ and $B = \max(5, \lceil 26 \ln(2SHK/\delta) \rceil)$, originating from the proven bound in Lemma 13 (Feige, 2004).
- **VIBE with Conjectured Constant:** Substitutes Feige's conjectured extremal constant $1/e$ into Lemma 13, setting $\alpha = 0.10$ and $B = \max(5, \lceil 2e \ln(2SHK/\delta) \rceil)$.

The conjectured constant reduces ensemble size B by $\approx 4.8\times$, while maintaining optimism at $\alpha = 0.10$ (see discussion after Theorem 4).

Empirical Results and Discussion. Across all evaluated domains (Figs. 1 to 3), *VIBE* with the conjectured constant consistently achieves the lowest or near-lowest cumulative regret, outperforming the leading baseline competitors—primarily *EQO* (Lee & Oh, 2025) and *ORLC* (Dann et al., 2019). Theoretical vanilla *VIBE* also exhibits sub-linear regret across all scales without tuning, though its larger ensemble size B incurs higher initial exploration.

On **RiverSwim** (Fig. 1), our results faithfully reproduce the findings of Lee & Oh (2025), where all algorithms were evaluated under theoretical closed-form parameters. In this domain, *VIBE* with the conjectured constant significantly outperforms *EQO* and *ORLC* across all four scaling dimensions.

On the **Gap MDP** (Fig. 2), evaluation code was not provided in the original release of Lee & Oh (2025), preventing exact reproduction; nevertheless, our evaluation obtains similar findings regarding the relative ranking of methods. Our modification incorporates awareness of state-dependent

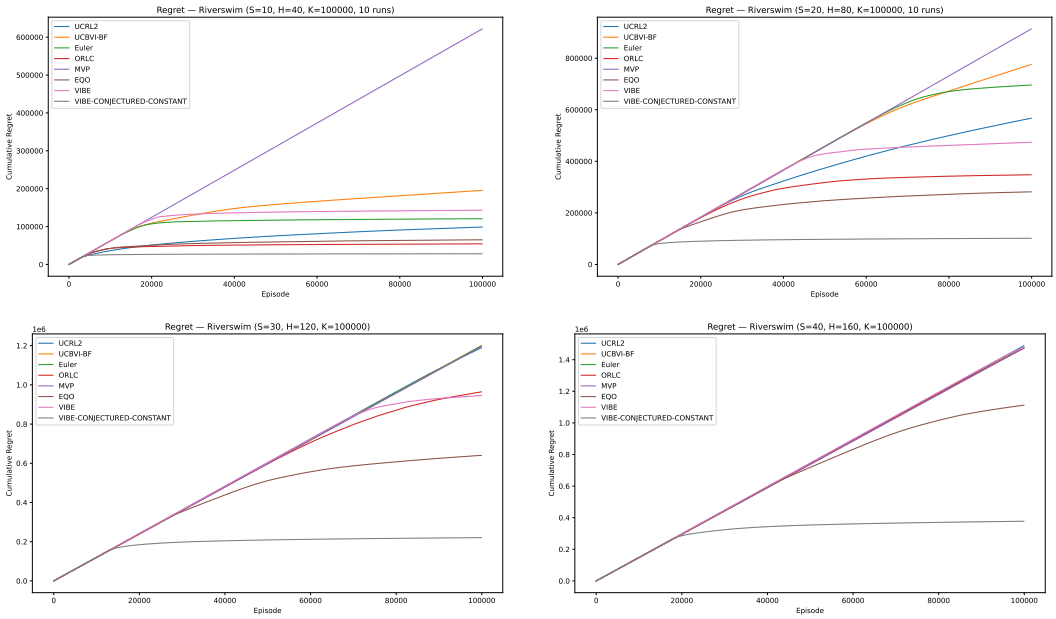


Figure 1: Cumulative regret on RiverSwim across four scaling configurations (mean across 10 independent runs).

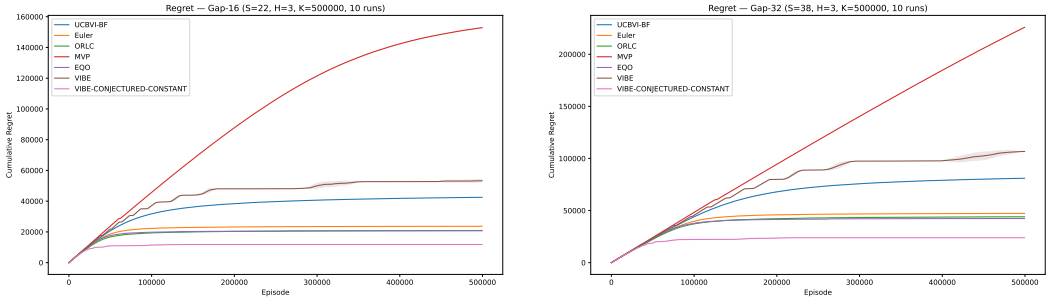


Figure 2: Cumulative regret on the Gap MDP ($n = 16, 32$) (mean across 10 independent runs).

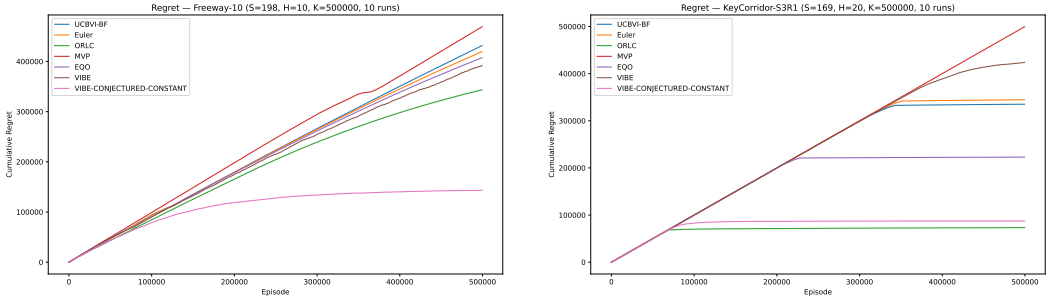


Figure 3: Cumulative regret on the BRIDGE benchmarks (mean across 10 independent runs).

action set sizes into the algorithms, which is undocumented as to whether it was previously included. In this benchmark, *VIBE* with the conjectured constant outperforms both *EQO* and *ORLC* across both $n = 16$ and $n = 32$, eliminating suboptimal branches more rapidly.

On the **BRIDGE** benchmarks (Fig. 3), code was likewise not provided by Lee & Oh (2025). Evaluating all algorithms strictly under their closed-form theoretical parameters, *VIBE* significantly out-

performs all baselines on `Freeway-10`. On `KeyCorridor-S3R1`, *VIBE* is nearly on par with *ORLC*, with both successfully reaching the sparse goal, whereas the remaining baselines fail to find the reward within the episode budget.

Overall, these evaluations show that *VIBE* performs well even in its vanilla formulation, and especially with the conjectured constant. Crucially, unlike all other methods, our exploration mechanism can be scaled heuristically, making it a particularly interesting approach for practical exploration.