ROBUST REINFORCEMENT LEARNING UNDER HEAVY-TAILED DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study reinforcement learning under heavy-tailed dynamics, where rare but large state transitions can make the Bellman target heavy-tailed and its variance need not be finite. In this setting, standard sample-mean-based Bellman updates can be sensitive to large observations and do not admit the usual light-tailed concentration guarantees. To address this issue, we develop a robust reinforcement learning algorithm based on the median-of-means (MoM) estimator. We incorporate MoM into the Bellman update and construct an exploration bonus from its high-probability estimation bound. A key challenge is that Bellman targets collected during learning are adaptive because the value estimates change over time. We handle this by comparing the adaptive targets with fixed reference Bellman targets and separating the statistical estimation error from the propagated value estimation error. Under moment and stability conditions, we establish an explicit sublinear expected regret bound for an unbounded state space. The bound depends on both the moment condition of the reference Bellman target and how strongly large state deviations can be controlled. Numerical results further show that MoM substantially reduces cumulative regret and more reliably selects the optimal action in a synthetic heavy-tailed environment.

1 INTRODUCTION

Reinforcement learning (RL) provides a general framework for sequential decision-making under uncertainty, where an agent learns a policy through interaction with an unknown environment (Sutton et al., 1998). Many stochastic systems exhibit heavy-tailed dynamics, where rare events generate extremely large state transitions or costs. Examples include bursty workloads in cloud systems and sudden traffic surges in communication networks. Such behavior can be modeled by heavy-tailed distributions whose variance may be infinite (Resnick, 2007).

Standard RL methods are not suitable for such environments. In particular, most RL algorithms rely on estimating the expected Bellman target using the sample mean (Azar et al., 2017). This approach works well when the variance is finite. However, under heavy-tailed distributions, a single large sample can significantly distort the estimate, leading to unstable learning and inaccurate value updates. Moreover, commonly used concentration tools, such as Hoeffding and Bernstein inequalities, do not apply when the variance is infinite (Catoni, 2012), making it difficult to obtain reliable performance guarantees. As a result, existing RL methods can suffer from unstable learning and poor performance in heavy-tailed environments.

In this work, we develop a robust RL framework based on the median-of-means (MoM) estimator (Nemirovskij & Yudin, 1983; Lu et al., 2019) for heavy-tailed Bellman targets. A central difficulty is that targets collected during learning are adaptive because the value estimates evolve over time. We address this by introducing fixed reference Bellman targets and decomposing the estimation error into a statistical MoM error and propagated value estimation errors. Under moment and stability conditions, we obtain a sublinear expected regret bound for an unbounded state space.

Our contributions can be summarized as follows:

- We formulate RL under heavy-tailed dynamics with an unbounded state space, where rare large transitions can invalidate standard light-tailed concentration arguments.

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.

- We develop a MoM-based Bellman estimator and RL algorithm, and handle adaptive targets through a decomposition using fixed reference Bellman targets.
- We establish an explicit sublinear expected regret bound for the unbounded state setting under moment and stability conditions, and numerically show lower cumulative regret, improved policy quality, and more reliable optimal-action selection in a synthetic heavy-tailed environment.

2 Related work

Heavy-tailed learning has been extensively studied in bandit settings, where robust estimators and truncation methods yield regret guarantees under heavy-tailed rewards (Bubeck et al., 2013; Lee & Lim, 2022; Agrawal et al., 2021; Lu et al., 2019; Xue et al., 2020; Medina & Yang, 2016; Xue et al., 2023; Zhao et al., 2025). However, these settings typically do not capture state-dependent dynamics. Several works extend robust estimation to RL, including heavy-tailed reward settings, temporal-difference learning, offline RL, and function approximation (Zhuang & Sui, 2021; Park & Lee, 2024; Cayci & Eryilmaz, 2023; Zhu et al., 2024; Huang et al., 2023). In contrast, we study heavy-tailed dynamics that affect both Bellman targets and state evolution in an unbounded state space. RL in unbounded state spaces has also been studied using stability analysis and Lyapunov functions (Mobin et al., 2014; Pavse et al., 2023; Shah et al., 2020), but these works generally rely on light-tailed dynamics and standard concentration tools, which is not suitable for heavy-tailed dynamics. Other robust decision-making methods address outliers, contamination, or heavy-tailed risk, while classical finite-horizon RL regret results typically assume bounded or sub-Gaussian settings (Kim & Lim, 2016; Wu et al., 2023; Simchi-Levi et al., 2022; Azar et al., 2017). In contrast, we jointly consider heavy-tailed dynamics, state-dependent decisions, and unbounded state evolution, and derive an explicit sublinear expected regret guarantee.

3 Problem formulation

We consider an episodic Markov decision process (MDP) with episodes $k = 1, 2, \ldots, K$ and steps in each episode $h = 1, \ldots, H$. At step (k, h), the system is in state $x_h^k \in \mathcal{X} \subseteq \mathbb{R}$ and chooses an action $a_h^k \in \mathcal{A} = \{0, 1, \ldots, \bar{a}\}$, where $\mathcal{X}$ is countably infinite and unbounded. The system evolves according to an unknown transition kernel $x_{h+1}^k \sim P(\cdot \mid x_h^k, a_h^k)$, where P is the transition probability. We assume polynomially heavy-tailed transition increments: for sufficiently large u,

$$c_1 u^{-\alpha} \leq \mathbb{P}(|x_{h+1}^k - x_h^k| > u \mid x_h^k, a_h^k) \leq c_2 u^{-\alpha}, \tag{1}$$

for some $0 < c_1 \leq c_2$ and tail index $\alpha \in (1, 2]$. That is, moments of order $r < \alpha$ are finite, while finite variance is not required. Additionally, we assume that the state evolves continuously across episodes, i.e., $x_1^{k+1} = x_{H+1}^k$ for $k = 1, \ldots, K-1$. At each step, the system incurs a cost $c(x_h^k, a_h^k) = g(x_h^k) + f(a_h^k)$, where g and f are nonnegative and g is increasing. A policy $\pi = (\pi_1, \ldots, \pi_H)$ specifies an action distribution $\pi_h(\cdot \mid x_h^k)$ at each state and step.

Our goal is to minimize the expected cumulative cost over a finite horizon. For a policy π, define

$$V_h^\pi(x) := \mathbb{E}\left[\sum_{t=h}^H c(x_t, a_t) \middle| x_h = x\right], \tag{2}$$

with $V_{H+1}^\pi = 0$, and let $V_h^\star(x) := \inf_\pi V_h^\pi(x)$. Over K episodes, let π^k denote the policy used by the learning algorithm in episode k, and let x_1^k denote the initial state of that episode. We define the regret as

$$\mathrm{Regret}(K) := \mathbb{E}\left[\sum_{k=1}^K \left(V_1^{\pi^k}(x_1^k) - V_1^\star(x_1^k)\right)\right], \tag{3}$$

where the expectation is taken over the randomness induced by the learning algorithm and the transition kernel.

Relevant examples. The model includes control systems with countable state spaces subject to heavy-tailed disturbances, such as $x_{h+1}^k = Ax_h^k + Ba_h^k + w_h^k$. In particular, $A = 1$ and $B = -1$ yields a queueing model in which the state is the accumulated workload and the action is the amount of service. Heavy-tailed disturbances produce unbounded state evolution and rare large deviations.

4 CHALLENGES IN HEAVY-TAILED REINFORCEMENT LEARNING

At step (k, h), the Bellman target is

$$Y_h^k := c(x_h^k, a_h^k) + V_{h+1}(x_{h+1}^k), \tag{4}$$

with conditional mean $\mu(x_h^k, a_h^k) := \mathbb{E}[Y_h^k \mid x_h^k, a_h^k]$. A standard approach is to estimate $\mu(x_h^k, a_h^k)$ using the sample mean of n independent observations. Under light-tailed assumptions, such as bounded or sub-Gaussian distributions, the sample mean satisfies an exponential concentration bound Vershynin (2020). Specifically, for i.i.d. samples $\{Y_i\}_{i=1}^n$, we have $\mathbb{P}\big(\big|\frac{1}{n}\sum_{i=1}^n Y_i - \mu\big| \geq t\big) \leq 2\exp(-cnt^2)$, for some constant $c > 0$. This inequality shows that the deviation probability decays exponentially in nt^2. Equivalently, for a fixed confidence level, the estimation error scales on the order of $\sqrt{1/n}$, leading to the standard $O(1/\sqrt{n})$ rate. However, under heavy-tailed dynamics, rare large state transitions can produce large costs and value estimates. As a result, the Bellman target can be heavy-tailed and may have infinite variance. Consequently, the sample mean is sensitive to rare extreme observations and standard sub-Gaussian concentration guarantees, including the $O(1/\sqrt{n})$ high-probability rate, no longer apply in general (Catoni, 2012). Under only a finite $(1+\epsilon)$-th moment, the sample mean can still converge to the true mean but can have poor high-probability behavior. When the estimation error bound is required to hold with probability at least $1-\delta$, the bound can become much larger as the failure probability δ decreases. This motivates the median-of-means estimator, which provides high-probability control under finite moment assumptions with only logarithmic dependence on $1/\delta$.

5 MOM-BASED BELLMAN ESTIMATION

5.1 MEDIAN-OF-MEANS ESTIMATOR

For a fixed step h, state-action pair (x, a), and value function V_{h+1}, suppose that we observe n i.i.d. Bellman targets $Y_1, \ldots, Y_n$. Partition the n samples into m disjoint, nearly equal blocks $\mathcal{B}_1, \ldots, \mathcal{B}_m$ with $b_j := |\mathcal{B}_j| \in \big\{\lfloor \frac{n}{m} \rfloor, \lceil \frac{n}{m} \rceil\big\}$. For each block $\mathcal{B}_j$, we compute the sample mean

$$\bar{Y}_j = \tfrac{1}{b_j}\textstyle\sum_{i\in\mathcal{B}_j} Y_i, \quad j = 1, \ldots, m. \tag{5}$$

The median-of-means estimator is then defined as

$$\hat{\mu}_{\mathrm{MoM}} = \mathrm{median}(\bar{Y}_1, \ldots, \bar{Y}_m). \tag{6}$$

Because the median is unaffected as long as a majority of block means remain accurate, MoM is less sensitive than the sample mean to rare extreme observations, which makes the MoM estimator robust to heavy-tailed samples. In the next subsection, we show that the MoM estimator admits a high-probability error bound under mild moment assumptions, which is used to construct the exploration bonus.

5.2 ERROR BOUND UNDER HEAVY-TAILED DYNAMICS

We assume that the Bellman target Y has finite $(1+\epsilon)$-th moment for some $\epsilon \in (0, 1)$:

$$\mathbb{E}\left[|Y - \mu|^{1+\epsilon}\right] \leq M_{1+\epsilon}, \qquad \mu = \mathbb{E}[Y]. \tag{7}$$

Note that this assumption is milder than the finite variance condition, which may fail under heavy-tailed dynamics. Under this mild moment assumption, the MoM estimator satisfies the following high-probability bound.

Lemma 1 (MoM Estimation Error Bound) *Let $Y_1, \ldots, Y_n$ be i.i.d. samples of a Bellman target Y with mean $\mu = \mathbb{E}[Y]$. Suppose that, for some $\epsilon \in (0, 1)$ and $M_{1+\epsilon} > 0$, $\mathbb{E}[|Y - \mu|^{1+\epsilon}] \leq M_{1+\epsilon}$. For a confidence level $\delta \in (0, 1)$, choose $m = \lceil c_0 \log(2/\delta) \rceil$ for a sufficiently large constant $c_0 > 0$, and assume $n \geq m$. Then, the median-of-means estimator satisfies*

$$|\hat{\mu}_{\mathrm{MoM}} - \mu| \leq C_\epsilon M_{1+\epsilon}^{\frac{1}{1+\epsilon}} \left(\tfrac{\log(2/\delta)}{n}\right)^{\frac{\epsilon}{1+\epsilon}} \tag{8}$$

with probability at least $1-\delta$, where $C_\epsilon > 0$ depends only on ϵ.

Proof sketch. Von Bahr-Esseen and Markov inequalities show that each block mean is accurate with a fixed positive probability. Then, a Chernoff bound implies that a majority of blocks are accurate, yielding the stated MoM bound. See Appendix A.

This lemma shows that the median-of-means estimator provides a stable estimate even when the variance of Y is infinite. Although the convergence rate is slower than the standard $O(1/\sqrt{n})$ rate achieved under sub-Gaussian (light-tailed) assumptions, it still yields a meaningful high-probability guarantee under heavy-tailed dynamics.

The exponent $\epsilon/(1+\epsilon)$ shows how the tail parameter affects the estimation rate. As $\epsilon \to 1$, the rate approaches $n^{-1/2}$. (i.e., the dependence of the error bound on the sample size approaches the standard $O(1/\sqrt{n})$ rate.) In contrast, when $\epsilon \to 0$, the moment condition becomes weaker, and the resulting error bound decays more slowly. To illustrate the dependence on ϵ and n, Fig. 1 plots the bound after removing the multiplicative factor $C_\epsilon M_{1+\epsilon}^{1/(1+\epsilon)}$.

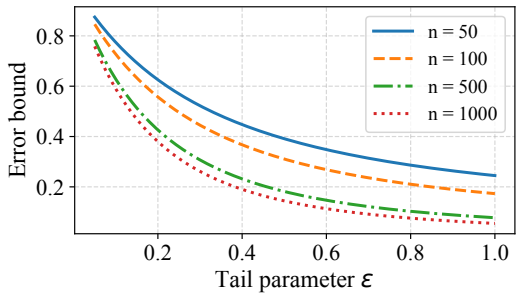


Figure 1: Normalized MoM error bound $(\log(2/\delta)/n)^{\epsilon/(1+\epsilon)}$ as a function of ϵ, with $\delta = 0.05$.

6 MoM-based reinforcement learning algorithm

In this section, we develop a reinforcement learning algorithm that incorporates the MoM estimator into the Bellman update. Let $Q_h(x,a)$ denote the action-value function, with $Q_{H+1} = 0$ and $Q_h(x,a) = 0$ for unvisited step-state-action tuples. Using $V_{h+1}(x) = \min_{a'} Q_{h+1}(x,a')$, the Bellman target is

$$Y_h^k = c(x_h^k, a_h^k) + \min_{a'} Q_{h+1}(x_{h+1}^k, a'). \tag{9}$$

For each tuple (h, x_h^k, a_h^k), we maintain m blocks of samples and assign incoming Bellman targets in round-robin order. Let $n = n_h(x_h^k, a_h^k)$ be the number of samples collected for (h, x_h^k, a_h^k), and let $\hat{\mu}_{\mathrm{MoM}}(x,a)$ be the MoM estimate as in equation 6. We update the action-value function as

$$Q_h^{\mathrm{new}}(x_h^k, a_h^k) \leftarrow (1-\eta_n) Q_h^{\mathrm{old}}(x_h^k, a_h^k) + \eta_n \left(\hat{\mu}_{\mathrm{MoM}}(x_h^k, a_h^k) - \beta_n(x_h^k, a_h^k) \right), \tag{10}$$

where $\eta_n \in (0,1]$ is the learning rate and, motivated by Lemma 1, we define the statistical exploration bonus as

$$\beta_n(x_h^k, a_h^k) = C_\epsilon M_{1+\epsilon}^{\frac{1}{1+\epsilon}} \left(\frac{\log(2/\delta)}{n} \right)^{\frac{\epsilon}{1+\epsilon}}. \tag{11}$$

Here, $M_{1+\epsilon}$ is a common moment bound. The action is selected according to the optimistic principle $a_h^k = \arg\min_{a \in \mathcal{A}} Q_h(x_h^k, a)$. The pseudocode is given in Appendix B. In standard RL, $\hat{\mu}_{\mathrm{MoM}}$ in equation 10 is replaced by the sample mean $\hat{\mu}_{\mathrm{mean}}(x_h^k, a_h^k) = \frac{1}{n}\sum_{i=1}^n Y_i$, and the exploration bonus is derived from light-tailed concentration bounds. Our use of MoM and the corresponding exploration bonus enables the regret analysis in the next section.

7 Regret analysis under heavy-tailed dynamics

Let $T = KH$, with H fixed. To handle the unbounded state space, we introduce a truncation threshold $U > 0$ on the state magnitude. This threshold is used only for analysis and does not modify the system dynamics or the algorithm. We decompose the regret into contributions from states satisfying $|x_h^k| \le U$ and $|x_h^k| > U$. We first present the final regret bound, then analyze adaptive Bellman estimation and the in-region and out-of-region contributions which are combined later to obtain the final regret bound.

7.1 FINAL REGRET BOUND

Theorem 1 (Regret Bound under Heavy-Tailed Dynamics) *Consider Algorithm 1 with fixed horizon H and $T = KH$. Suppose that the state space growth condition equation 21, the Lipschitz cost condition equation 22, and the uniform increment moment bound equation 23 hold for some $r \in (1, \alpha)$. Assume that the moment condition for the reference Bellman targets in Lemma 2 holds uniformly over all (h, x, a), for some $\epsilon \in (0, 1)$ and a common bound $M_{1+\epsilon}$. Also, assume that the state process generated during learning satisfies the drift condition equation 28, with $W(x) = |x|^r$ and $\gamma = r - 1$. The initial state x_1^1 is drawn from a fixed distribution satisfying $\mathbb{E}[|x_1^1|^r] < \infty$. All constants in these conditions are independent of T and U.*

Use a fixed learning rate $\eta_n = \eta_0 \in (0, 1]$, set $\delta = (T + 1)^{-4}$, and choose $m = \lceil c_0 \log(2/\delta) \rceil$ as in Lemma 1. Use the common moment bound $M_{1+\epsilon}$ in the exploration bonus. Then, the expected regret defined in equation 3 satisfies

$$\mathrm{Regret}(K) \le \widetilde{O}\Big((KH)^{1 - \frac{\epsilon\gamma}{(1+\epsilon)(1+\gamma)}}\Big). \tag{12}$$

Here, $\widetilde{O}(\cdot)$ hides polylogarithmic factors in $KH + 1$. The hidden constants may depend on the fixed horizon, learning rate, model and drift parameters, and initial moment, but not on K.

The bound is sublinear for any fixed $\epsilon > 0$ and $\gamma > 0$. A smaller ϵ gives a weaker moment condition on the reference Bellman target, while a smaller γ gives weaker control of large state deviations. Both lead to a larger regret exponent. Under an extension to light-tailed dynamics admitting arbitrarily high-order moment and drift control, the same exponent approaches $1/2$ as $\epsilon \to 1$ and $\gamma \to \infty$, recovering the form of the standard $\widetilde{O}(\sqrt{KH})$ rate.

Proof sketch. Lemma 2 separates the adaptive Bellman estimation error into a reference MoM error and propagated next-step value errors. Using optimism, this yields the in-region bound in Lemma 3. The drift and increment moment conditions control the out-of-region and boundary terms in Lemma 4. Combining the two bounds gives the result of

$$\mathrm{Regret}(K) \le \widetilde{O}\Big(M_{1+\epsilon}^{\frac{1}{1+\epsilon}}(1+U)^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}} + (1+U)^2 + TU^{-\gamma}\Big). \tag{13}$$

Then, choosing the optimized $U = T^{\epsilon/((1+\epsilon)(1+\gamma))}$ that balances the dominant terms yields the final bound equation 12.

7.2 ADAPTIVE BELLMAN ESTIMATION ANALYSIS

Lemma 1 provides a high-probability concentration bound for i.i.d. samples from a fixed distribution. However, Bellman targets generated during learning are adaptive because the value estimates change over time, and thus are generally not identically distributed. We therefore decompose each target into a fixed reference Bellman target and an adaptive correction term.

Fix a step h and a state-action pair (x, a). Let i index the visits to this fixed tuple (h, x, a), rather than the episodes. Let X_i' denote the next state observed at the i-th visit, and let V_{h+1}^i denote the value estimate used to construct the target at that visit. The adaptive and reference Bellman targets are defined as $Y_i = c(x, a) + V_{h+1}^i(X_i')$ and $Y_i^\star = c(x, a) + V_{h+1}^\star(X_i')$, respectively. Unlike V_{h+1}^i, $V_{h+1}^\star$ is fixed throughout learning, and the reference targets are introduced only for analysis and are not computed by the algorithm. We define the mean of the reference target under the transition kernel $P(\cdot \mid x, a)$ as $\mu_h^\star(x, a) := c(x, a) + \mathbb{E}_{X' \sim P(\cdot|x,a)}\left[V_{h+1}^\star(X')\right]$. Here, X' denotes a next state drawn from $P(\cdot \mid x, a)$. When this expectation is finite, the Bellman optimality equation gives $\mu_h^\star(x, a) = Q_h^\star(x, a)$. To describe the adaptive correction, define $\Delta_{i,h+1}(x') := V_{h+1}^i(x') - V_{h+1}^\star(x')$. Then, $Y_i = Y_i^\star + \Delta_{i,h+1}(X_i')$. The following lemma bounds the adaptive estimation error by a statistical estimation error and a correction term from the next-state value estimates.

Lemma 2 (Adaptive Bellman Error Decomposition) *Fix (h, x, a) and a confidence parameter $\delta \in (0, 1)$. Choose $m = \lceil c_0 \log(2/\delta) \rceil$ as in Lemma 1, and let $n \ge m$ be a fixed sample size. Suppose that the reference target has finite mean $\mu_h^\star(x, a)$ and satisfies $\mathbb{E}\left[|Y_i^\star - \mu_h^\star(x, a)|^{1+\epsilon}\right] \le M_{1+\epsilon}$ for some $\epsilon \in (0, 1)$ and $0 < M_{1+\epsilon} < \infty$. Also, assume that the observed adaptive targets are finite almost surely. Use the same round-robin partition $\mathcal{B}_1, \ldots, \mathcal{B}_m$ for the first n adaptive and reference*

targets, with $b_j = |\mathcal{B}_j|$ as defined in Section 5.1. Then, with probability at least $1-\delta$, the following bound holds whenever the first n visits to (h,x,a) have occurred:

$$|\hat{\mu}_{\mathrm{MoM}} - \mu_h^\star(x,a)| \le \beta_n(x,a) + \max_{1\le j\le m} \left| \tfrac{1}{b_j} \textstyle\sum_{i\in\mathcal{B}_j} \Delta_{i,h+1}(X_i') \right|, \tag{14}$$

where $\beta_n(x,a) = C_\epsilon M_{1+\epsilon}^{\frac{1}{1+\epsilon}} \left(\frac{\log(2/\delta)}{n}\right)^{\frac{\epsilon}{1+\epsilon}}$ is the statistical estimation error bound in Lemma 1. Here, $\hat{\mu}_{\mathrm{MoM}}$ is computed from these first n adaptive targets.

The first term controls the reference estimation error, while the second term captures the largest absolute block average of the next-step value errors from past visits.

Proof sketch. For a fixed (h,x,a), the reference targets are i.i.d., so Lemma 1 controls their MoM estimation error. Each adaptive block mean equals the corresponding reference block mean plus the average adaptive correction, and the difference between the two medians is bounded by the largest absolute average correction within a block. The triangle inequality yields the stated bound. The detailed proof is provided in Appendix C. This lemma gives a decomposition for a fixed tuple and a fixed sample size. The simultaneous confidence argument over multiple tuples and sample sizes is handled in the in-region regret analysis.

7.3 In-region regret bound

To relate the Bellman estimation error to the regret defined in equation 3, we define the optimal action gap as

$$A_h^\star(x,a) := Q_h^\star(x,a) - V_h^\star(x). \tag{15}$$

Since $V_h^\star(x) = \min_{a\in\mathcal{A}} Q_h^\star(x,a)$, we have $A_h^\star(x,a) \ge 0$. Using the Bellman equations, for any policy π,

$$V_h^\pi(x) - V_h^\star(x) = \mathbb{E}_\pi\Big[A_h^\star(x_h,a_h) + V_{h+1}^\pi(x_{h+1}) - V_{h+1}^\star(x_{h+1}) \mid x_h = x\Big]. \tag{16}$$

Applying this relation recursively over the finite horizon and using $V_{H+1}^\pi = V_{H+1}^\star = 0$, we obtain

$$V_1^\pi(x_1) - V_1^\star(x_1) = \mathbb{E}_\pi\left[\textstyle\sum_{h=1}^H A_h^\star(x_h,a_h) \mid x_1\right]. \tag{17}$$

Therefore, the regret in equation 3 can equivalently be written as

$$\mathrm{Regret}(K) = \mathbb{E}\left[\textstyle\sum_{k=1}^K \sum_{h=1}^H A_h^\star(x_h^k,a_h^k)\right]. \tag{18}$$

For a threshold $U>0$, we define

$$\mathrm{Regret}_{\mathrm{in}}(K;U) := \mathbb{E}\left[\textstyle\sum_{k=1}^K \sum_{h=1}^H A_h^\star(x_h^k,a_h^k)\mathbf{1}\{|x_h^k| \le U\}\right], \tag{19}$$

where $\mathbf{1}$ is the indicator function. We then define $\mathrm{Regret}_{\mathrm{out}}(K;U)$ in the same way with the indicator $\mathbf{1}\{|x_h^k| > U\}$. Hence,

$$\mathrm{Regret}(K) = \mathrm{Regret}_{\mathrm{in}}(K;U) + \mathrm{Regret}_{\mathrm{out}}(K;U). \tag{20}$$

In this subsection, we focus on the in-region regret.

For the following analysis, we impose additional conditions on the state space, cost, and transitions. First, assume

$$|\mathcal{X} \cap [-u,u]| \le C_{\mathcal{X}}(1+u), \qquad u>0. \tag{21}$$

$C_{\mathcal{X}} > 0$ is a constant independent of u. This condition allows the number of states in the bounded region to be controlled by the truncation threshold. We also assume that, for some constant $L_g > 0$, g is Lipschitz as

$$|g(x) - g(y)| \le L_g|x-y|, \qquad x,y\in\mathcal{X}, \tag{22}$$

and that, for some $r \in (1,\alpha)$, the transition increments satisfy

$$\sup_{x\in\mathcal{X},\, a\in\mathcal{A}} \mathbb{E}_{X'\sim P(\cdot|x,a)}\left[|X'-x|^r\right] \le M_r < \infty. \tag{23}$$

Also, the moment condition in Lemma 1 for reference targets is assumed to hold uniformly over all (h, x, a) with the same bound $M_{1+\epsilon}$.

We use a constant learning rate $\eta_n = \eta_0 \in (0, 1]$, set $\delta = (T+1)^{-4}$, and choose $m = \lceil c_0 \log(2/\delta) \rceil$ as in Lemma 1. These choices ensure that the reference estimation bounds hold simultaneously throughout learning with high probability.

Let Q_h^k denote the estimate used to select the action at step (k, h). Consider the high-probability event on which the reference MoM bounds used throughout the analysis hold simultaneously. On this event, zero initialization and the optimistic update preserve $Q_h^k(x, a) \le Q_h^\star(x, a)$. Because the action is selected greedily, we have

$$A_h^\star(x_h^k, a_h^k) \le Q_h^\star(x_h^k, a_h^k) - Q_h^k(x_h^k, a_h^k). \tag{24}$$

Thus, the in-region regret is controlled by the corresponding Q-estimation errors.

Further, the next state may lie outside the bounded region even when the current state lies inside it. To include this contribution, the analysis uses the auxiliary regions

$$\mathcal{X}_h(U) := \{x \in \mathcal{X} : |x| \le hU\}, \qquad h = 1, \dots, H. \tag{25}$$

These regions are introduced only for the analysis of the recursive errors. Then, define the boundary contribution as

$$\mathcal{B}(K; U) := \mathbb{E}[\textstyle\sum_{k=1}^{K} \sum_{h=1}^{H-1} (1 + |x_{h+1}^k|) \times \mathbf{1}\{|x_h^k| \le hU,\ |x_{h+1}^k| > (h+1)U\}]. \tag{26}$$

This term accounts for transitions from $\mathcal{X}_h(U)$ to states outside $\mathcal{X}_{h+1}(U)$. For such transitions, the next-step value estimation error cannot be propagated using the recursion within the bounded region. Under the Lipschitz cost condition and the uniform transition moment bound, the optimal value grows at most linearly with the state magnitude. Therefore, the remaining next-step value error can be bounded by a constant multiple of $1 + |x_{h+1}^k|$.

Lemma 3 (In-Region Regret Bound) *Under the conditions and parameter choices stated above, for any $U > 0$, the expected in-region regret satisfies*

$$\mathrm{Regret}_{\mathrm{in}}(K; U) \le \widetilde{O}\left(M_{1+\epsilon}^{\frac{1}{1+\epsilon}} (1+U)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}} + (1+U)^2 + \mathcal{B}(K; U) \right). \tag{27}$$

Here, $\widetilde{O}(\cdot)$ hides polylogarithmic factors in $T + 1$. The hidden constants may depend on the fixed horizon, learning rate, and model parameters, but not on T or U.

The first term is the statistical learning error obtained by summing the MoM error bounds over the visited tuples. The second term accounts for zero initialization and the visits before all m blocks are nonempty. The third term accounts for the next-step errors outside the auxiliary regions. It will be bounded together with the out-of-region contribution in the next subsection.

Proof sketch. Using the independent reference sequences in Appendix C, we construct a simultaneous confidence event over the visited tuples and their sample counts. Since there are at most T visited tuples and T possible sample counts, the choice $\delta = (T+1)^{-4}$ gives an overall failure probability of at most T^{-2}. On this event, the monotonicity of the MoM estimator and the subtraction of β_n preserve optimism. Applying Lemma 2 and summing the updates with a constant learning rate over the visits to each tuple decomposes the cumulative estimation error into statistical, initialization, and propagated next-step errors. The next-step errors are propagated over the fixed horizon, while the state space growth condition and the linear growth of the optimal values control the statistical and initialization terms. Transitions leaving the auxiliary regions contribute the boundary term $\mathcal{B}(K; U)$. Finally, the uniformly bounded optimal action gaps make the contribution of the confidence failure event $O(T^{-1})$, which is absorbed into the stated bound. The detailed proof is provided in Appendix D.

7.4 Out-of-region regret bound

We next bound the out-of-region regret and the boundary contribution $\mathcal{B}(K; U)$ defined in equation 26. The out-of-region regret depends on how often the state exceeds U, whereas the boundary contribution depends on transitions that leave the auxiliary regions. To control the state process during learning, we impose an additional stability condition.

Let $r \in (1, \alpha)$ be the moment order in equation 23, and set $\gamma := r - 1 > 0$. We use the Lyapunov function $W(x) := |x|^r$. Assume that there exist finite constants $\kappa > 0$ and $B_W \geq 0$ such that, at every step (k, h) during learning,

$$\mathbb{E}\left[W(x_{h+1}^k) - W(x_h^k) \mid x_h^k, a_h^k\right] \leq B_W - \kappa |x_h^k|^\gamma. \tag{28}$$

The constants are independent of k, h, T, and U. The inequality in equation 28 is the stability condition assumed to hold at every step of the learning process. We impose equation 28 as an additional assumption for the out-of-region analysis. We also assume a fixed initial-state distribution with $\mathbb{E}[W(x_1^1)] < \infty$. Under these assumptions and the conditions in Section 7.3, we obtain the following result.

Lemma 4 (Out-of-Region and Boundary Bounds) *Under the conditions in Section 7.3, the drift condition equation 28, and the assumption on the initial moment above, for any $U > 0$,*

$$\mathrm{Regret}_{\mathrm{out}}(K; U) + \mathcal{B}(K; U) \leq CTU^{-\gamma}, \qquad \gamma = r - 1. \tag{29}$$

Here, $C > 0$ is independent of T and U but may depend on the fixed horizon, model and drift parameters, and $\mathbb{E}[W(x_1^1)]$.

Proof sketch. Taking expectations in equation 28 and summing over all $T = KH$ steps, the intermediate Lyapunov terms cancel because $x_1^{k+1} = x_{H+1}^k$. Since $W \geq 0$, we obtain

$$\textstyle\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}[|x_h^k|^\gamma] \leq \frac{B_W T + \mathbb{E}[W(x_1^1)]}{\kappa}. \tag{30}$$

Applying Markov's inequality and the uniform action-gap bound $A_h^\star(x, a) \leq G_H$ in equation 63 gives

$$\mathrm{Regret}_{\mathrm{out}}(K; U) \leq G_H \textstyle\sum_{k=1}^K \sum_{h=1}^H \mathbb{P}(|x_h^k| > U) \leq CTU^{-\gamma}. \tag{31}$$

Thus, the out-of-region regret is controlled by the number of out-of-region visits and the bounded action gaps. For the boundary contribution, a transition satisfying $|x_h^k| \leq hU$ and $|x_{h+1}^k| > (h+1)U$ must also satisfy $|x_{h+1}^k - x_h^k| > U$. When $U \geq 1$, its weight satisfies

$$1 + |x_{h+1}^k| \leq (H+1)|x_{h+1}^k - x_h^k|. \tag{32}$$

The uniform increment moment bound gives

$$\mathbb{E}_{X' \sim P(\cdot|x,a)}\left[|X' - x| \mathbf{1}\{|X' - x| > U\}\right] \leq M_r U^{-(r-1)}. \tag{33}$$

Summing over the boundary transitions therefore yields $\mathcal{B}(K; U) \leq CTU^{-\gamma}$ for $U \geq 1$. For $0 < U < 1$, the current state in a boundary transition satisfies $|x_h^k| \leq H$. The uniform increment moment bound then gives $\mathcal{B}(K; U) \leq CT \leq CTU^{-\gamma}$. Combining these bounds proves the lemma above. The detailed proof is provided in Appendix E.

7.5 Combining the bounds

We now combine the in-region and out-of-region bounds under the conditions and parameter choices in Sections 7.3 and 7.4. Using equation 20 and Lemmas 3 and 4, for any $U > 0$, we obtain

$$\mathrm{Regret}(K) \leq \widetilde{O}\Big(M_{1+\epsilon}^{\frac{1}{1+\epsilon}} (1+U)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}} + (1+U)^2 + TU^{-\gamma}\Big). \tag{34}$$

The first two terms represent the statistical and initialization errors. The last term bounds both the out-of-region regret and the boundary contribution. These terms give a trade-off in the choice of U. Choosing a larger U increases the first two terms but reduces the last term, while a smaller U has the opposite effect. To obtain a simple regret bound, we choose

$$U = T^{\frac{\epsilon}{(1+\epsilon)(1+\gamma)}}. \tag{35}$$

This gives

$$\mathrm{Regret}(K) \leq \widetilde{O}\Big((KH)^{1 - \frac{\epsilon\gamma}{(1+\epsilon)(1+\gamma)}}\Big). \tag{36}$$

$\widetilde{O}(\cdot)$ hides polylogarithmic factors in $T + 1$. The hidden constants may depend on the fixed horizon, learning rate, model and drift parameters, and initial moment, but not on T. Because $\epsilon > 0$ and $\gamma > 0$, this gives a sublinear expected regret bound.

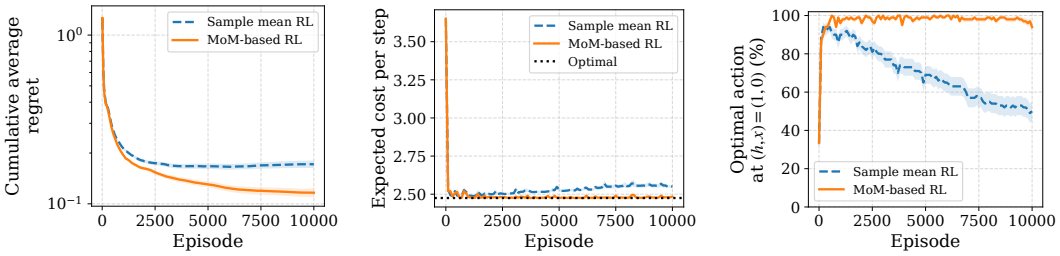
Proof sketch. Write $\theta = \epsilon/(1+\epsilon)$. The threshold in equation 35 satisfies $1 \leq U \leq T^{1-\theta}$. Thus, $(1+U)^\theta \leq 2^\theta U$ and $(1+U)^2 \leq 4UT^{1-\theta}$. Since $M_{1+\epsilon}$ is fixed, the upper bound in equation 34 is therefore at most $\widetilde{O}(UT^{1-\theta} + TU^{-\gamma})$ at this choice of U. The relation $U^{1+\gamma} = T^\theta$ makes these two terms equal, yielding equation 36. The detailed derivation is provided in Appendix F.

8 NUMERICAL RESULTS

We compare the sample-mean and MoM-based RL methods in a synthetic heavy-tailed MDP designed to expose the learning algorithms to rare but large state transitions. The environment contains three actions with different costs and exposure probabilities to large workload states. The burst distribution has an unbounded Pareto component with tail index $\alpha = 1.5$. We use $H = 2$, $K = 10000$, and 100 independent runs. Both methods use the same initialization, greedy action selection, learning rate, and block structure. The bonus coefficient is selected separately for each method using an independent set of tuning runs. The detailed setup is provided in Appendix G.

Fig. 2(a) shows the cumulative regret divided by the number of elapsed steps. At the end of learning, the average values are 0.172 for the sample-mean method and 0.116 for MoM, corresponding to a 32.2% reduction in cumulative regret. This indicates that the MoM-based method incurs less loss from suboptimal action selections over the same training budget. Fig. 2(b) evaluates the quality of the learned policy at each checkpoint. We fix the current Q-function, form its greedy policy, and compute the expected cumulative cost starting from the same initial state $x = 0$; the value is divided by H and reported as the expected cost per step. The MoM-based method remains close to the optimal value, whereas the sample-mean method shows a larger gap. At the final checkpoint, the expected costs per step are approximately 2.484 for MoM and 2.550 for the sample mean, compared with the optimal value of 2.475. Fig. 2(c) shows the fraction of the 100 independent runs in which the greedy policy selects the optimal action at $(h, x) = (1, 0)$. At the final checkpoint, MoM selects the optimal action in 94 runs, compared with 50 runs for the sample-mean method.

These results show a consistent difference in the learning behavior of the two estimators in this heavy-tailed environment. Rare large Bellman targets can substantially affect a sample mean and change the ranking of the actions, whereas the MoM estimator is less sensitive to such observations. As a result, the MoM-based method selects the optimal action more reliably, learns a policy whose expected cost remains closer to the optimal value, and accumulates substantially less regret over the same finite training budget. These results support the use of robust Bellman estimation when rare large observations can strongly affect value estimates.



(a) Cumulative average regret (b) Expected policy cost (c) Optimal-action rate

Figure 2: Performance comparison in a synthetic heavy-tailed environment. Curves show averages over 100 independent runs, with shaded regions representing one standard error around the mean.

9 CONCLUSION

In this work, we studied reinforcement learning under heavy-tailed dynamics with an unbounded state space. Rare but large state transitions can make Bellman targets heavy-tailed, making standard sample-mean-based updates sensitive to large observations. To address this issue, we developed a MoM-based reinforcement learning algorithm with an exploration bonus derived from a robust estimation bound. We also handled the adaptive Bellman targets generated during learning by comparing them with fixed reference targets and accounting for propagated value estimation errors. Under moment and stability conditions, we established an explicit sublinear expected regret bound. The bound captures the effect of both the Bellman target moment condition and large state deviations. Our numerical results further showed that MoM achieves substantially lower cumulative regret, learns a policy closer to the optimal one, and selects the optimal action more reliably in a synthetic heavy-tailed environment. These results support robust Bellman estimation as a useful approach for reducing the effect of rare large observations on reinforcement learning.

AI USE STATEMENT

Generative AI was used to assist with proof direction and verification, research ideation, manuscript drafting and language editing, paper organization, and the design, implementation, and parameter tuning of numerical experiments. All mathematical arguments developed with AI assistance, simulation code and settings, and written content were reviewed and verified by the authors.

ETHICS STATEMENT

This work is based on theoretical analysis and simulations. It does not involve human subjects, personal data, or real-world deployment. We are not aware of any specific ethical concerns associated with this work.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs of the theoretical results are provided in Appendices A– F, including detailed pseudocode for the proposed MoM-based RL algorithm in Appendix B. Appendix G provides the complete simulation setting, including the workload model, learning parameters, bonus calibration, evaluation procedure, oracle computation, and estimation experiment used in the numerical results. These details are provided to support reproduction of both the theoretical and numerical results.

ACKNOWLEDGMENTS

Will be added later.

REFERENCES

Shubhada Agrawal, Sandeep K Juneja, and Wouter M Koolen. Regret minimization in heavy-tailed bandits. In *Conference on Learning Theory*, pp. 26–62. PMLR, 2021.

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.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pp. 1148–1185, 2012.

Semih Cayci and Atilla Eryilmaz. Provably robust temporal difference learning for heavy-tailed rewards. *Advances in Neural Information Processing Systems*, 36:25693–25711, 2023.

Jiayi Huang, Han Zhong, Liwei Wang, and Lin Yang. Tackling heavy-tailed rewards in reinforcement learning with function approximation: Minimax optimal and instance-dependent regret bounds. *Advances in Neural Information Processing Systems*, 36:56576–56588, 2023.

Michael Jong Kim and Andrew EB Lim. Robust multiarmed bandit problems. *Management Science*, 62(1):264–285, 2016.

Kyungjae Lee and Sungbin Lim. Minimax optimal bandits for heavy tail rewards. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5280–5294, 2022.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In *international conference on machine learning*, pp. 4154–4163. PMLR, 2019.

Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pp. 1642–1650. PMLR, 2016.

Shariq A Mobin, James A Arnemann, and Friedrich T Sommer. Information-based learning by agents in unbounded state spaces. *Advances in Neural Information Processing Systems*, 27, 2014.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Hyeon-Jun Park and Kyungjae Lee. Heavy-tailed reinforcement learning with penalized robust estimator. *IEEE Access*, 12:107800–107817, 2024.

Brahma S Pavse, Matthew Zurek, Yudong Chen, Qiaomin Xie, and Josiah P Hanna. Learning to stabilize online reinforcement learning in unbounded state spaces. *arXiv preprint arXiv:2306.01896*, 2023.

Sidney I Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer, 2007.

Devavrat Shah, Qiaomin Xie, and Zhi Xu. Stable reinforcement learning with unbounded state space. *arXiv preprint arXiv:2006.04353*, 2020.

David Simchi-Levi, Zeyu Zheng, and Feng Zhu. A simple and optimal policy design for online learning with safety against heavy-tailed risk. *Advances in Neural Information Processing Systems*, 35:33795–33805, 2022.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Roman Vershynin. High-dimensional probability. *University of California, Irvine*, 10(11):31, 2020.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, 1 r 2. *The Annals of Mathematical Statistics*, pp. 299–303, 1965.

Yulian Wu, Xingyu Zhou, Youming Tao, and Di Wang. On private and robust bandits. *Advances in Neural Information Processing Systems*, 36:34778–34790, 2023.

Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. *arXiv preprint arXiv:2004.13465*, 2020.

Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. *Advances in Neural Information Processing Systems*, 36: 70880–70891, 2023.

Canzhe Zhao, Shinji Ito, and Shuai Li. Heavy-tailed linear bandits: Adversarial robustness, best-of-both-worlds, and beyond. *arXiv preprint arXiv:2508.13679*, 2025.

Jin Zhu, Runzhe Wan, Zhengling Qi, Shikai Luo, and Chengchun Shi. Robust offline reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pp. 541–549. PMLR, 2024.

Vincent Zhuang and Yanan Sui. No-regret reinforcement learning with heavy-tailed rewards. pp. 3385–3393, 2021.

A PROOF OF THE MOM ESTIMATION ERROR BOUND

We analyze the deviation of the median-of-means estimator introduced in Section 5.1 to prove Lemma 1. We first study the deviation of a single block mean. Let $p = 1 + \epsilon \in (1, 2)$, and define $X_i = Y_i - \mu$, where $\{Y_i\}$ are i.i.d. samples of the fixed Bellman target distribution considered in Lemma 1 and $\mu = \mathbb{E}[Y]$. For convenience, let $M_p := \mathbb{E}[|Y - \mu|^p]$, where $M_p \leq M_{1+\epsilon}$ by assumption. Let $b_{\min} := \min_{1 \leq j \leq m} b_j = \lfloor \frac{n}{m} \rfloor$. Then, $\bar{Y}_j - \mu = \frac{1}{b_j} \sum_{i \in \mathcal{B}_j} X_i$. Hence,

$$\mathbb{E}\left[|\bar{Y}_j - \mu|^p\right] = \tfrac{1}{b_j^p} \mathbb{E}\left[\left|\textstyle\sum_{i \in \mathcal{B}_j} X_i\right|^p\right]. \tag{37}$$

Since $p \in (1, 2)$, we apply the von Bahr–Esseen inequality von Bahr & Esseen (1965) to obtain

$$\mathbb{E}\left[\left|\textstyle\sum_{i \in \mathcal{B}_j} X_i\right|^p\right] \leq C_p \textstyle\sum_{i \in \mathcal{B}_j} \mathbb{E}[|X_i|^p], \tag{38}$$

where $C_p > 0$ depends only on p, and the inequality holds for independent mean-zero random variables. This inequality is suitable for heavy-tailed random variables with finite p-th moments. Since the samples are i.i.d., all X_i have the same distribution, and thus we have

$$\textstyle\sum_{i \in \mathcal{B}_j} \mathbb{E}[|X_i|^p] = b_j \, \mathbb{E}[|X_1|^p] = b_j \, \mathbb{E}[|Y - \mu|^p] = b_j M_p. \tag{39}$$

Therefore,

$$\mathbb{E}\left[|\bar{Y}_j - \mu|^p\right] \leq \tfrac{C_p M_p}{b_j^{p-1}} \leq \tfrac{C_p M_p}{b_{\min}^{p-1}}. \tag{40}$$

We next convert this moment bound into a tail bound. By Markov's inequality, for any $t > 0$,

$$\mathbb{P}\left(|\bar{Y}_j - \mu| \geq t\right) \leq \tfrac{\mathbb{E}[|\bar{Y}_j - \mu|^p]}{t^p} \leq \tfrac{C_p M_p}{b_{\min}^{p-1} t^p}. \tag{41}$$

By choosing $t = (4 C_p M_p)^{1/p} b_{\min}^{-(p-1)/p}$, we obtain

$$\mathbb{P}\left(|\bar{Y}_j - \mu| \leq t\right) \geq \tfrac{3}{4}. \tag{42}$$

Define the indicator variables

$$Z_j := \mathbf{1}\{|\bar{Y}_j - \mu| \leq t\}, \quad j = 1, \ldots, m. \tag{43}$$

Then $\{Z_j\}$ are independent Bernoulli random variables with $\mathbb{E}[Z_j] \geq \frac{3}{4}$ since the blocks are constructed from disjoint subsets of i.i.d. samples. This representation allows us to treat each block as a Bernoulli trial indicating whether it is a "good" estimate, which enables the use of concentration bounds for sums of independent random variables. Let $S = \sum_{j=1}^{m} Z_j$ denote the number of "good" blocks. Then $\mathbb{E}[S] \geq \frac{3m}{4}$.

We are interested in the event that fewer than half of the blocks are good, i.e., $S \leq m/2$. Using a Chernoff bound for sums of independent Bernoulli variables, we obtain

$$\mathbb{P}(S \leq m/2) \leq \exp(-cm), \tag{44}$$

for some constant $c > 0$.

Whenever $S > m/2$, more than half of the block means lie within t of μ, which implies that the median also lies within this range. Therefore,

$$\mathbb{P}\left(|\hat{\mu}_{\mathrm{MoM}} - \mu| > t\right) \leq \exp(-cm). \tag{45}$$

To express the bound in terms of a target confidence level δ, we choose $m = \left\lceil c_0 \log\left(\frac{2}{\delta}\right) \right\rceil$ for a sufficiently large constant $c_0 > 0$ such that $\exp(-cm) \leq \delta$. Since $n \geq m$, every block is nonempty. Recalling that $t = (4 C_p M_p)^{1/p} b_{\min}^{-(p-1)/p}$ and substituting $b_{\min} = \lfloor n/m \rfloor$ and $p = 1 + \epsilon$, we obtain

$$t \leq C_\epsilon M_{1+\epsilon}^{\frac{1}{1+\epsilon}} \left(\tfrac{\log(2/\delta)}{n}\right)^{\frac{\epsilon}{1+\epsilon}},$$

where $C_\epsilon > 0$ depends only on ϵ. This completes the proof of Lemma 1.

B MoM-BASED REINFORCEMENT LEARNING ALGORITHM

For completeness, we provide a detailed pseudocode description of the proposed MoM-based reinforcement learning algorithm.

Algorithm 1 MoM-based Reinforcement Learning

Set $Q_{H+1}(x,a)=0$ for all (x,a), and use $Q_h(x,a)=0$ as the default value for any unvisited (h,x,a), $h=1,\dots,H$.
For each (h,x,a), initialize its Q-function entry and m sample buffers when the tuple is first visited.
for episode $k=1,\dots,K$ **do**
  Observe initial state x_1^k
  for $h=1,\dots,H$ **do**
    Select action $a_h^k=\arg\min_{a\in\mathcal{A}} Q_h(x_h^k,a)$
    Observe cost $c(x_h^k,a_h^k)$ and next state x_{h+1}^k
    If (h,x_h^k,a_h^k) is visited for the first time, create its m empty sample buffers
    Compute sample $Y_h^k=c(x_h^k,a_h^k)+\min_{a'} Q_{h+1}(x_{h+1}^k,a')$
    Assign Y_h^k to one of the m blocks of (h,x_h^k,a_h^k) in a round-robin manner
    Update the corresponding block mean
    Let $n=n_h(x_h^k,a_h^k)$ be the number of samples collected for (h,x_h^k,a_h^k)
    if $n\ge m$ **then**
      Compute MoM estimate $\hat{\mu}_{\mathrm{MoM}}=\mathrm{median}(\bar{Y}_1,\dots,\bar{Y}_m)$
      Compute bonus $\beta_n=C_\epsilon M^{\frac{1}{1+\epsilon}}\left(\frac{\log(2/\delta)}{n}\right)^{\frac{\epsilon}{1+\epsilon}}$
      Update $Q_h^{\mathrm{new}}(x_h^k,a_h^k)\leftarrow(1-\eta_n)Q_h^{\mathrm{old}}(x_h^k,a_h^k)+\eta_n(\hat{\mu}_{\mathrm{MoM}}-\beta_n)$
    end if
  end for
end for

The system evolves continuously across episodes. In particular, the initial state of episode $k+1$ is given by $x_1^{k+1}=x_{H+1}^k$. For a newly visited step-state-action tuple, the algorithm first collects samples until all m blocks are nonempty. The MoM estimate and the corresponding Q-value remain at its initial value of zero.

C PROOF OF THE ADAPTIVE BELLMAN ERROR DECOMPOSITION

We prove Lemma 2. The main difficulty is that the Bellman targets generated during learning use value estimates that change over time. Therefore, the adaptive targets cannot be directly treated as i.i.d. samples from a fixed distribution. We first introduce reference targets constructed using the fixed optimal value function. We then compare the adaptive targets with these reference targets.

Fix a step h, a state-action pair (x,a), and a confidence parameter $\delta\in(0,1)$. Choose $m=\lceil c_0\log(2/\delta)\rceil$ as in Lemma 1, and fix an integer $n\ge m$. The n is chosen before observing the samples. It is not the random number of visits made by a particular point in the learning process. Let $\mathcal{R}_n$ denote the event that (h,x,a) is visited at least n times during learning. The comparison with the adaptive estimator will be made on $\mathcal{R}_n$, when all of its first n targets are available.

We first explain how the next-state samples can be represented. For each step-state-action tuple, imagine an independent sequence of next-state draws from its transition kernel. These sequences are independent of the randomization used by the learning algorithm. For the fixed tuple (h,x,a), denote its sequence by $X_1',X_2',\dots$, where $X_i'\sim P(\cdot\mid x,a)$ independently across $i=1,2,\dots$. At the i-th visit to (h,x,a), the next unused draw X_i' is revealed as the next state.

We define the reference Bellman targets for this sequence as

$$Y_i^\star:=c(x,a)+V_{h+1}^\star(X_i'). \tag{46}$$

For the fixed tuple (h,x,a), both $c(x,a)$ and $V_{h+1}^\star$ are fixed. Therefore, the reference targets $Y_1^\star,Y_2^\star,\dots$ are i.i.d. samples from a fixed distribution.

Let

$$\mu_h^\star(x,a) := \mathbb{E}[Y_i^\star] = c(x,a) + \mathbb{E}_{X'\sim P(\cdot|x,a)}\left[V_{h+1}^\star(X')\right]. \tag{47}$$

Here, X' denotes a next state drawn from $P(\cdot \mid x,a)$. As assumed in Lemma 2, this mean is finite, and the reference targets satisfy

$$\mathbb{E}\left[|Y_i^\star - \mu_h^\star(x,a)|^{1+\epsilon}\right] \le M_{1+\epsilon} \qquad \epsilon \in (0,1). \tag{48}$$

This is the moment condition used when applying Lemma 1 to the reference targets.

On the event $\mathcal{R}_n$, the first n visits to (h,x,a) occur. For $i = 1,\dots,n$, let V_{h+1}^i denote the value estimate used to construct the Bellman target at the i-th visit. The adaptive target is

$$Y_i = c(x,a) + V_{h+1}^i(X_i'). \tag{49}$$

As assumed in Lemma 2, the observed adaptive targets are finite almost surely. Note that V_{h+1}^i depends on the previous learning history, and this is why the adaptive targets are not necessarily i.i.d.

Define the next-step value error at an observed visit as

$$\Delta_{i,h+1}(x') := V_{h+1}^i(x') - V_{h+1}^\star(x'). \tag{50}$$

The two targets use the same immediate cost and the same next state, so

$$Y_i = Y_i^\star + \Delta_{i,h+1}(X_i'). \tag{51}$$

We next consider the block structure. Partition the indices $\{1,\dots,n\}$ into $\mathcal{B}_1,\dots,\mathcal{B}_m$ using the round-robin rule in Algorithm 1. The block assignment depends only on the visit index. Using the same notation in Section 5.1, we have $b_j = |\mathcal{B}_j|$ and $b_j \in \left\{\left\lfloor \frac{n}{m}\right\rfloor, \left\lceil \frac{n}{m}\right\rceil\right\}$. Since $n \ge m$, every block is nonempty. We use exactly the same index partition for the adaptive and reference targets.

For each block, define the reference block mean as

$$\bar{Y}_j^\star := \tfrac{1}{b_j}\textstyle\sum_{i\in\mathcal{B}_j} Y_i^\star. \tag{52}$$

On $\mathcal{R}_n$, the adaptive block means are also available and are given by

$$\bar{Y}_j := \tfrac{1}{b_j}\textstyle\sum_{i\in\mathcal{B}_j} Y_i. \tag{53}$$

Using equation 51, on $\mathcal{R}_n$ we have

$$\bar{Y}_j = \bar{Y}_j^\star + \tfrac{1}{b_j}\textstyle\sum_{i\in\mathcal{B}_j} \Delta_{i,h+1}(X_i'). \tag{54}$$

Let

$$\hat{\mu}_{\mathrm{MoM}}^\star := \mathrm{median}\left(\bar{Y}_1^\star,\dots,\bar{Y}_m^\star\right) \tag{55}$$

denote the reference MoM estimator. The adaptive estimator $\hat{\mu}_{\mathrm{MoM}}$ uses the same median convention. We assume that, when m is even, we define the median as the average of the two middle values after sorting the block means.

The first n reference targets are i.i.d., satisfy the moment condition equation 48, and are partitioned into the nonempty blocks specified above. Therefore, Lemma 1 gives

$$|\hat{\mu}_{\mathrm{MoM}}^\star - \mu_h^\star(x,a)| \le \beta_n(x,a), \tag{56}$$

with probability at least $1-\delta$, where $\beta_n(x,a) = C_\epsilon M_{1+\epsilon}^{\frac{1}{1+\epsilon}}\left(\frac{\log(2/\delta)}{n}\right)^{\frac{\epsilon}{1+\epsilon}}$.

Define the reference confidence event as

$$\mathcal{E}_n := \left\{|\hat{\mu}_{\mathrm{MoM}}^\star - \mu_h^\star(x,a)| \le \beta_n(x,a)\right\}. \tag{57}$$

The argument above establishes $\mathbb{P}(\mathcal{E}_n^c) \le \delta$.

We now compare the adaptive and reference MoM estimators. For any two collections of real numbers $\{u_j\}_{j=1}^m$ and $\{v_j\}_{j=1}^m$, the median satisfies

$$|\mathrm{median}(u_1,\dots,u_m) - \mathrm{median}(v_1,\dots,v_m)| \le \max_{1\le j\le m} |u_j - v_j|. \tag{58}$$

Indeed, if $|u_j - v_j| \le r$ for every j, then each coordinate changes by at most r, and therefore the corresponding median also changes by at most r.

Applying equation 58 to the block decomposition in equation 54, we obtain, on $\mathcal{R}_n$,

$$|\hat{\mu}_{\mathrm{MoM}} - \hat{\mu}^\star_{\mathrm{MoM}}| \le \max_{1\le j\le m} \left| \tfrac{1}{b_j} \textstyle\sum_{i\in\mathcal{B}_j} \Delta_{i,h+1}(X'_i) \right|. \tag{59}$$

Thus, the difference between the two estimators is bounded by the largest absolute average correction within a block. Combining this inequality with the reference bound in equation 56, on $\mathcal{R}_n \cap \mathcal{E}_n$ we have

$$\begin{aligned} |\hat{\mu}_{\mathrm{MoM}} - \mu^\star_h(x,a)| &\le |\hat{\mu}_{\mathrm{MoM}} - \hat{\mu}^\star_{\mathrm{MoM}}| + |\hat{\mu}^\star_{\mathrm{MoM}} - \mu^\star_h(x,a)| \\ &\le \beta_n(x,a) + \max_{1\le j\le m} \left| \tfrac{1}{b_j} \textstyle\sum_{i\in\mathcal{B}_j} \Delta_{i,h+1}(X'_i) \right|. \end{aligned} \tag{60}$$

The bound in equation 60 holds on $\mathcal{R}_n \cap \mathcal{E}_n$. Since $\mathbb{P}\left(\mathcal{R}_n \cap \mathcal{E}_n^c\right) \le \mathbb{P}(\mathcal{E}_n^c) \le \delta$, the probability that the first n visits occur and this bound fails is at most δ. This completes the proof of Lemma 2.

D Proof of the in-region regret bound

We prove Lemma 3 under the conditions and parameter choices stated in Section 7.3. Fix $U > 0$, let $T = KH$, and write $\theta := \frac{\epsilon}{1+\epsilon},\ L := \log(2/\delta),\ \delta = (T+1)^{-4}$. Then, $1 - \theta = 1/(1+\epsilon)$ and $\beta_n = C_\epsilon M_{1+\epsilon}^{1/(1+\epsilon)} L^\theta n^{-\theta}$. The bonus is common to all tuples because the same reference moment bound is assumed. Throughout the proof, $C > 0$ denotes a generic constant that may change from line to line, and all such constants are independent of T and U. Sums over empty index sets are taken to be zero.

Value and action-gap bounds. Let $M_1 := M_r^{1/r}$ and $F_{\max} := \max_{a\in\mathcal{A}} f(a) < \infty$. The uniform increment moment condition implies $\mathbb{E}[|X' - x| \mid x, a] \le M_1$. By summing the increments, under any policy and conditional on $x_h = x$,

$$\mathbb{E}_\pi[|x_{h+j} - x| \mid x_h = x] \le jM_1. \tag{61}$$

Let $\ell = H - h + 1$ and $D_\ell = L_g M_1 \ell(\ell-1)/2$. The Lipschitz cost condition and $0 \le f(a) \le F_{\max}$ give

$$\ell g(x) - D_\ell \le V_h^\pi(x) \le \ell g(x) + D_\ell + \ell F_{\max}. \tag{62}$$

The same bounds apply when the first action is fixed, and hence apply to both $V_h^\star(x)$ and $Q_h^\star(x,a)$. Therefore,

$$0 \le A_h^\star(x,a) \le 2D_\ell + \ell F_{\max} \le G_H, \qquad G_H := L_g M_1 H(H-1) + HF_{\max}. \tag{63}$$

In particular, the action gaps are bounded even though the costs need not be bounded. Choose any fixed $x_{\mathrm{ref}} \in \mathcal{X}$. Because $g(x) \le g(x_{\mathrm{ref}}) + L_g(|x| + |x_{\mathrm{ref}}|)$, equation 62 also gives a constant $C_V > 0$ such that

$$0 \le V_h^\star(x) \le Q_h^\star(x,a) \le C_V(1 + |x|). \tag{64}$$

A simultaneous reference confidence event. Use the independent next-state sequences introduced in Appendix C to define the reference targets. Let $\tau_1, \tau_2, \ldots$ denote the distinct step-state-action tuples in the order of their first visits. Consider a tuple $\tau_j = (h_j, x_j, a_j)$ at the time of its first visit. Conditional on the learning history up to this point, the tuple τ_j is fixed, and the next-state samples used to construct its reference targets are i.i.d. draws from $P(\cdot \mid x_j, a_j)$. Since the same moment bound for the reference targets holds for every tuple, Lemma 1 can be applied to these reference targets. For each fixed sample size $n \in \{m, \ldots, T\}$, the failure probability is therefore at most δ.

Let $\mathcal{G}$ be the event that the MoM concentration bound for the reference targets holds for every distinct tuple visited during learning and every sample size $n \in \{m, \ldots, T\}$. Since the reference targets are defined for all indices of the underlying next-state samples, this event is well defined even for samples sizes that are not reached by the actual visits.

During T steps, at most T distinct tuples can be visited, and there are at most T possible samples sizes. Therefore, applying a union bound to the conditional failure probability above gives

$$\mathbb{P}(\mathcal{G}^c) \le T^2\delta = \tfrac{T^2}{(T+1)^4}. \tag{65}$$

This argument does not require the set of visited tuples to be fixed or the corresponding error events to be independent. If $m > T$, no MoM update occurs, so $\mathcal{G}$ has no constraints.

Optimism and a lower bound on the estimates. We first work on $\mathcal{G}$. Initially, $Q_h(x,a) = 0 \le Q_h^\star(x,a)$ because the costs are nonnegative. Suppose optimism has held before the current update. Every stored target then satisfies $Y_i \le Y_i^\star$. The same inequality holds for the corresponding block means and, by monotonicity of the median, $\hat{\mu}_{\mathrm{MoM}} \le \hat{\mu}^\star_{\mathrm{MoM}}$. Consequently,

$$\hat{\mu}_{\mathrm{MoM}} - \beta_n \le \hat{\mu}^\star_{\mathrm{MoM}} - \beta_n \le Q_h^\star(x,a). \tag{66}$$

The update is a convex combination of this quantity and the previous optimistic estimate. Tuples with fewer than m samples remain at zero. Induction over the updates therefore gives

$$Q_h^k(x,a) \le Q_h^\star(x,a) \tag{67}$$

throughout learning on $\mathcal{G}$.

We also need a lower bound on the learned values. For every update, $n \ge m \ge c_0 L$, so

$$\beta_n \le B_0, \qquad B_0 := C_\epsilon M_{1+\epsilon}^{1/(1+\epsilon)} c_0^{-\theta}. \tag{68}$$

The constant B_0 does not depend on T. If the next-step estimates are bounded below by $-(H-h)B_0$, then every target at step h is bounded below by that quantity. Subtracting the bonus gives a lower bound of $-(H-h+1)B_0$, which is preserved by the convex-combination update. Starting from the terminal value zero, backward induction yields

$$Q_h^k(x,a) \ge -(H-h+1)B_0 \tag{69}$$

for every learning history, without requiring $\mathcal{G}$.

Let $V_h^k(x) := \min_a Q_h^k(x,a)$, with the estimates taken just before action selection at step (k,h). Combining equation 64, equation 67, and equation 69, on $\mathcal{G}$ there is a constant C_d such that

$$0 \le V_h^\star(x) - V_h^k(x) \le C_d(1+|x|). \tag{70}$$

Moreover, greedy action selection gives

$$Q_h^k(x_h^k, a_h^k) = \min_a Q_h^k(x_h^k, a) \le \min_a Q_h^\star(x_h^k, a) = V_h^\star(x_h^k). \tag{71}$$

Thus,

$$A_h^\star(x_h^k, a_h^k) \le Q_h^\star(x_h^k, a_h^k) - Q_h^k(x_h^k, a_h^k). \tag{72}$$

Summing the updates for one tuple. Fix (h,x,a) and let $N = N_h(x,a)$ be its total number of visits. Let $Q_{h,n}(x,a)$ be its estimate after processing the n-th sample, and define

$$e_{h,n} := Q_h^\star(x,a) - Q_{h,n}(x,a). \tag{73}$$

At the n-th visit, the action is selected using $Q_{h,n-1}$. Also, $e_{h,n} = Q_h^\star(x,a)$ for $0 \le n < m$. On $\mathcal{G}$, all these errors are nonnegative. For each actual visit i, write

$$d_i := V_{h+1}^\star(X_i') - V_{h+1}^i(X_i') = -\Delta_{i,h+1}(X_i') \ge 0. \tag{74}$$

For $h = H$, we have $d_i = 0$ by the terminal condition.

Let $\hat{\mu}_{h,n}$ be the MoM estimate from the first n stored targets of this tuple. Lemma 2 and $b_j \ge \lfloor n/m \rfloor \ge n/(2m)$ give, for $n \ge m$,

$$\begin{aligned} Q_h^\star(x,a) - \hat{\mu}_{h,n} &\le \beta_n + \max_j \tfrac{1}{b_j} \textstyle\sum_{i \in \mathcal{B}_j} d_i \\ &\le \beta_n + \tfrac{2m}{n} \textstyle\sum_{i=1}^n d_i. \end{aligned} \tag{75}$$

The block partition here is the partition after n samples. Using the constant learning rate, the update therefore satisfies

$$e_{h,n} \le (1-\eta_0) e_{h,n-1} + \eta_0 \left(2\beta_n + \tfrac{2m}{n} \textstyle\sum_{i=1}^n d_i\right). \tag{76}$$

After rearranging,

$$e_{h,n-1} \le 2\beta_n + \tfrac{2m}{n} \textstyle\sum_{i=1}^n d_i + \tfrac{e_{h,n-1} - e_{h,n}}{\eta_0}. \tag{77}$$

When we sum over $n = m, \dots, N$, the last terms cancel except for the boundary terms. Since $e_{h,m-1} = Q_h^\star(x,a)$ and $e_{h,N} \geq 0$, including the initial visits gives

$$\sum_{n=1}^{N} e_{h,n-1} \leq \left(m - 1 + \tfrac{1}{\eta_0}\right) Q_h^\star(x,a) + 2\sum_{n=m}^{N} \beta_n + 2m \sum_{n=m}^{N} \tfrac{1}{n} \sum_{i=1}^{n} d_i. \tag{78}$$

If $N < m$, no Q-update is performed for this tuple. In this case, the two sums over $n = m, \dots, N$ are zero, and the same upper bound remains valid.

The last term accounts for the repeated use of stored targets. Changing the order of summation, for $N \geq m$,

$$\sum_{n=m}^{N} \tfrac{1}{n} \sum_{i=1}^{n} d_i = \sum_{i=1}^{N} d_i \sum_{n=\max\{m,i\}}^{N} \tfrac{1}{n} \leq (1 + \log(T+1)) \sum_{i=1}^{N} d_i. \tag{79}$$

Set $\Lambda_T := 2m(1 + \log(T+1))$. Then, combining this with equation 78, each tuple's cumulative error is bounded by its initialization term, its statistical error sum, and $\Lambda_T \sum_{i=1}^{N} d_i$.

Propagation through the auxiliary regions. Using $\mathcal{X}_h(U)$ from equation 25, define

$$S_h : \sum_{k=1}^{K} \big(Q_h^\star(x_h^k, a_h^k) - Q_h^k(x_h^k, a_h^k)\big) \mathbf{1}\{|x_h^k| \leq hU\}, \qquad S_{H+1} := 0. \tag{80}$$

These sums are nonnegative on $\mathcal{G}$. Define the initialization and statistical terms at step h by

$$I_h := \left(m - 1 + \tfrac{1}{\eta_0}\right) \sum_{x \in \mathcal{X}_h(U),\, a \in \mathcal{A}} Q_h^\star(x,a), \tag{81}$$

$$J_h := 2 \sum_{x \in \mathcal{X}_h(U),\, a \in \mathcal{A}} \sum_{n=m}^{N_h(x,a)} \beta_n. \tag{82}$$

For $h < H$, let

$$Z_h := \sum_{k=1}^{K} (1 + |x_{h+1}^k|) \mathbf{1}\{|x_h^k| \leq hU,\ |x_{h+1}^k| > (h+1)U\}, \tag{83}$$

and set $Z_H := 0$. By definition, $\mathbb{E}[\sum_{h=1}^{H} Z_h] = \mathcal{B}(K; U)$.

Consider a sample collected at step $h < H$ in episode k. The update at this step changes only Q_h, not Q_{h+1}. Thus, the next-step value estimate stored in this target is the one used to select the action at step $(k, h+1)$. If $|x_{h+1}^k| \leq (h+1)U$, its value error is bounded by

$$V_{h+1}^\star(x_{h+1}^k) - V_{h+1}^k(x_{h+1}^k) \leq Q_{h+1}^\star(x_{h+1}^k, a_{h+1}^k) - Q_{h+1}^k(x_{h+1}^k, a_{h+1}^k). \tag{84}$$

Each sample corresponds to a distinct subsequent visit at step $h+1$. Consequently, the sum of these interior next-step errors is at most S_{h+1}. For transitions satisfying $|x_{h+1}^k| > (h+1)U$, the corresponding next-step value errors are bounded by equation 70. Therefore, their contribution is at most $C_d Z_h$. Summing equation 78 over the tuples in $\mathcal{X}_h(U) \times \mathcal{A}$ therefore yields

$$S_h \leq I_h + J_h + \Lambda_T S_{h+1} + C_d \Lambda_T Z_h. \tag{85}$$

We next bound $\sum_h I_h$ and $\sum_h J_h$. Let

$$D_U := \sum_{h=1}^{H} |\mathcal{X}_h(U)|\, |\mathcal{A}| \tag{86}$$

be the total number of tuples in the auxiliary regions. By equation 21 and fixed H, $D_U \leq C_D(1+U)$ for a constant C_D. Using equation 64, for the initialization terms, we obtain

$$\sum_{h=1}^{H} I_h \leq C \left(m + \tfrac{1}{\eta_0}\right) (1+U)^2. \tag{87}$$

For the statistical terms, the inequality $\sum_{n=1}^{N} n^{-\theta} \leq N^{1-\theta}/(1-\theta)$ gives

$$\sum_{h=1}^{H} J_h \leq C M_{1+\epsilon}^{1/(1+\epsilon)} L^\theta \sum_{h=1}^{H} \sum_{x \in \mathcal{X}_h(U),\, a \in \mathcal{A}} N_h(x,a)^{1-\theta}. \tag{88}$$

The total number of these visits is at most T. Since the function $f(z) = z^{1-\theta}$ is concave, Jensen's inequality gives

$$\sum_{h=1}^{H} \sum_{x \in \mathcal{X}_h(U),\, a \in \mathcal{A}} N_h(x,a)^{1-\theta} \leq D_U^\theta T^{1-\theta}. \tag{89}$$

Therefore,

$$\sum_{h=1}^{H} J_h \leq C M_{1+\epsilon}^{1/(1+\epsilon)} L^\theta (1+U)^\theta T^{1-\theta}. \tag{90}$$

Starting from $S_{H+1} = 0$, repeated application of equation 85 gives

$$S_h \le \sum_{j=h}^{H} \Lambda_T^{j-h}(I_j + J_j) + C_d \sum_{j=h}^{H-1} \Lambda_T^{j-h+1} Z_j. \tag{91}$$

Combining this with equation 87 and equation 90, on $\mathcal{G}$ we have

$$\sum_{h=1}^{H} S_h \le C(1+\Lambda_T)^H \Big[M_{1+\epsilon}^{1/(1+\epsilon)} L^\theta (1+U)^\theta T^{1-\theta} + \Big(m + \tfrac{1}{\eta_0}\Big)(1+U)^2 + \sum_{h=1}^{H} Z_h \Big]. \tag{92}$$

Here, $L = O(\log(T+1))$, $m = O(\log(T+1))$, and $\Lambda_T = O((1+\log(T+1))^2)$. Because H is fixed, the factor $(1+\Lambda_T)^H$ and the other terms involving L and m contribute only polylogarithmic factors in $T+1$.

Taking expectations. Let

$$R_{\mathrm{in}}^{\mathrm{path}} := \sum_{k=1}^{K} \sum_{h=1}^{H} A_h^\star(x_h^k, a_h^k) \mathbf{1}\{|x_h^k| \le U\}. \tag{93}$$

Then, $\mathrm{Regret}_{\mathrm{in}}(K;U) = \mathbb{E}[R_{\mathrm{in}}^{\mathrm{path}}]$ by definition. We decompose this expectation according to the confidence event $\mathcal{G}$:

$$\mathrm{Regret}_{\mathrm{in}}(K;U) = \mathbb{E}\left[R_{\mathrm{in}}^{\mathrm{path}} \mathbf{1}_{\mathcal{G}}\right] + \mathbb{E}\left[R_{\mathrm{in}}^{\mathrm{path}} \mathbf{1}_{\mathcal{G}^c}\right]. \tag{94}$$

On $\mathcal{G}$, equation 72 and $\{|x| \le U\} \subseteq \mathcal{X}_h(U)$ imply

$$R_{\mathrm{in}}^{\mathrm{path}} \le \sum_{h=1}^{H} S_h. \tag{95}$$

Taking expectations in equation 92, we use

$$\mathbb{E}\left[\mathbf{1}_{\mathcal{G}} \sum_h Z_h\right] \le \mathbb{E}\left[\sum_h Z_h\right] = \mathcal{B}(K;U). \tag{96}$$

This inequality follows from $Z_h \ge 0$ and $\mathbf{1}_{\mathcal{G}} \le 1$. Therefore,

$$\mathbb{E}\left[R_{\mathrm{in}}^{\mathrm{path}} \mathbf{1}_{\mathcal{G}}\right] \le \widetilde{O}\Big(M_{1+\epsilon}^{\frac{1}{1+\epsilon}} (1+U)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}} + (1+U)^2 + \mathcal{B}(K;U) \Big). \tag{97}$$

On the failure event $\mathcal{G}^c$, the uniform action-gap bound equation 63 gives

$$R_{\mathrm{in}}^{\mathrm{path}} \le G_H T. \tag{98}$$

Thus,

$$\mathbb{E}\left[R_{\mathrm{in}}^{\mathrm{path}} \mathbf{1}_{\mathcal{G}^c}\right] \le G_H T\, \mathbb{P}(\mathcal{G}^c) \le \tfrac{G_H T^3}{(T+1)^4} \le \tfrac{G_H}{T+1}. \tag{99}$$

Since $(1+U)^2 \ge 1$, the contribution from the failure event can be absorbed into the initialization term in equation 87.

Combining equation 94, equation 97, and equation 99, we obtain

$$\mathrm{Regret}_{\mathrm{in}}(K;U) \le \widetilde{O}\Big(M_{1+\epsilon}^{\frac{1}{1+\epsilon}} (1+U)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}} + (1+U)^2 + \mathcal{B}(K;U) \Big). \tag{100}$$

Here, $\widetilde{O}(\cdot)$ hides polylogarithmic factors in $T+1$. This completes the proof of Lemma 3.

E Proof of the out-of-region and boundary bounds

We prove Lemma 4 under the conditions in Section 7.4. Fix $U > 0$, let $T = KH$, and recall that $W(x) = |x|^r$ and $\gamma = r - 1 > 0$. Throughout the proof, $C > 0$ denotes a generic constant that may change from line to line and is independent of T and U.

Cumulative bound on the state moment. We first check that the expectations of W are finite at each step. Since $r > 1$, using $x' = x+(x'-x)$ and the standard inequality $|a+b|^r \le 2^{r-1}(|a|^r+|b|^r)$ for $r \ge 1$, we have

$$|x'|^r \le 2^{r-1}\big(|x|^r + |x'-x|^r\big). \tag{101}$$

Together with equation 23, this gives

$$\mathbb{E}[W(x_{h+1}^k)] \le 2^{r-1}\big(\mathbb{E}[W(x_h^k)] + M_r\big). \tag{102}$$

Starting from $\mathbb{E}[W(x_1^1)] < \infty$, the expectations are therefore finite at every finite step. Taking expectations in equation 28, we obtain

$$\mathbb{E}[W(x_{h+1}^k)] - \mathbb{E}[W(x_h^k)] \le B_W - \kappa\mathbb{E}[|x_h^k|^\gamma]. \tag{103}$$

Because $x_1^{k+1} = x_{H+1}^k$, the intermediate Lyapunov terms cancel when we sum over all T steps. Rearranging the resulting inequality gives

$$\kappa\textstyle\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}[|x_h^k|^\gamma] \le B_W T + \mathbb{E}[W(x_1^1)] - \mathbb{E}[W(x_{H+1}^K)] \le B_W T + \mathbb{E}[W(x_1^1)], \tag{104}$$

where the second inequality uses $W \ge 0$. Consequently,

$$\textstyle\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}[|x_h^k|^\gamma] \le \frac{B_W T+\mathbb{E}[W(x_1^1)]}{\kappa} \le CT. \tag{105}$$

The last inequality uses $T \ge 1$ and the fixed finite initial moment.

Out-of-region regret. We use the sum of the state moments to bound the expected number of visits outside $[-U, U]$. Applying Markov's inequality to the nonnegative random variable $|x_h^k|^\gamma$ gives

$$\mathbb{P}(|x_h^k| > U) = \mathbb{P}(|x_h^k|^\gamma > U^\gamma) \le \tfrac{\mathbb{E}[|x_h^k|^\gamma]}{U^\gamma}. \tag{106}$$

The uniform action-gap bound equation 63 holds for all states and actions. Thus, using the definition of the out-of-region regret,

$$\begin{aligned}\mathrm{Regret}_{\mathrm{out}}(K;U) &= \mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^H A_h^\star(x_h^k,a_h^k)\mathbf{1}\{|x_h^k|>U\}\right]\\ &\le G_H\sum_{k=1}^K\sum_{h=1}^H \mathbb{P}(|x_h^k|>U) \le \frac{G_H}{U^\gamma}\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}[|x_h^k|^\gamma] \le CTU^{-\gamma},\end{aligned} \tag{107}$$

where the last inequality follows from equation 105.

Boundary contribution. We next bound $\mathcal{B}(K;U)$ from equation 26 using the increment moment condition. Consider a transition with $|x_h^k| \le hU$ and $|x_{h+1}^k| > (h+1)U$, where $h < H$. By the triangle inequality,

$$|x_{h+1}^k - x_h^k| \ge |x_{h+1}^k| - |x_h^k| > U. \tag{108}$$

Thus, leaving the next auxiliary region requires an increment whose magnitude exceeds U.

First, suppose $U \ge 1$. For a boundary transition,

$$1 + |x_{h+1}^k| \le 1 + hU + |x_{h+1}^k - x_h^k| \le (H+1)|x_{h+1}^k - x_h^k|. \tag{109}$$

The second inequality uses $h \le H-1$, $U \ge 1$, and $|x_{h+1}^k - x_h^k| > U$. For any fixed (x,a), the uniform increment moment condition gives

$$\mathbb{E}_{X'\sim P(\cdot|x,a)}\left[|X'-x|\mathbf{1}\{|X'-x|>U\}\right] \le U^{-(r-1)}\mathbb{E}_{X'\sim P(\cdot|x,a)}[|X'-x|^r] \le M_r U^{-\gamma}. \tag{110}$$

On the event $|X'-x| > U$, we have $|X'-x|^{r-1} \ge U^{r-1}$ since $r > 1$. So, here, we used $|X'-x|\mathbf{1}\{|X'-x|>U\} \le U^{-(r-1)}|X'-x|^r$. Applying equation 110 conditional on x_h^k, a_h^k and then taking expectations, we obtain

$$\begin{aligned}\mathcal{B}(K;U) &\le (H+1)\textstyle\sum_{k=1}^K\sum_{h=1}^{H-1}\mathbb{E}\left[|x_{h+1}^k - x_h^k|\mathbf{1}\{|x_{h+1}^k - x_h^k|>U\}\right]\\ &\le (H+1)M_r T U^{-\gamma}, \qquad U \ge 1.\end{aligned} \tag{111}$$

The first inequality uses equation 108 and equation 109, and the second inequality uses $K(H-1) \leq T$.

For $0 < U < 1$, the current state in a boundary transition satisfies $|x_h^k| \leq hU \leq H$. Therefore, on these transitions,

$$1 + |x_{h+1}^k| \leq 1 + H + |x_{h+1}^k - x_h^k|. \tag{112}$$

Because $r > 1$, the uniform increment moment bound also implies $\mathbb{E}[|x_{h+1}^k - x_h^k|] \leq M_r^{1/r}$. Using the definition of $\mathcal{B}(K;U)$ and $\mathbf{1}\{\cdot\} \leq 1$, we have

$$\mathcal{B}(K;U) \leq T\big(1 + H + M_r^{1/r}\big) \leq CTU^{-\gamma}, \qquad 0 < U < 1, \tag{113}$$

since $U^{-\gamma} \geq 1$ in this case. Together, equation 111 and equation 113 show that $\mathcal{B}(K;U) \leq CTU^{-\gamma}$ for every $U > 0$.

Adding this boundary bound to equation 107, we conclude that

$$\mathrm{Regret}_{\mathrm{out}}(K;U) + \mathcal{B}(K;U) \leq CTU^{-\gamma}, \qquad U > 0. \tag{114}$$

The constant C may depend on H, G_H, M_r, κ, B_W, and $\mathbb{E}[W(x_1^1)]$, but not on T or U. This completes the proof of Lemma 4.

F Proof of the regret bound under heavy-tailed dynamics

We derive the combined regret bound in Section 7.5 under the conditions and parameter choices of Lemmas 3 and 4. Let $T = KH$ and $\theta := \epsilon/(1+\epsilon)$, so that $0 < \theta < 1/2$ and $1 - \theta = 1/(1+\epsilon)$. Throughout this proof, $\widetilde{O}(\cdot)$ hides polylogarithmic factors in $T+1$, with constants independent of T and U.

Lemma 4 bounds the sum of $\mathrm{Regret}_{\mathrm{out}}(K;U)$ and $\mathcal{B}(K;U)$ by a constant multiple of $TU^{-\gamma}$. Since both quantities are nonnegative, the same upper bound applies to each of them. Combining this result with Lemma 3 and equation 20, for any $U > 0$, we have

$$\mathrm{Regret}(K) \leq \widetilde{O}\Big(M_{1+\epsilon}^{\frac{1}{1+\epsilon}}(1+U)^{\theta}T^{1-\theta} + (1+U)^2 + TU^{-\gamma}\Big). \tag{115}$$

The polylogarithmic factor multiplying $\mathcal{B}(K;U)$ in the in-region bound remains included in $\widetilde{O}(\cdot)$. The moment bound $M_{1+\epsilon}$ is independent of T and U, so its factor can be included in the hidden constants.

We first simplify this upper bound for $1 \leq U \leq T^{1-\theta}$. Since $0 < \theta < 1$, we have

$$(1+U)^{\theta}T^{1-\theta} \leq 2^{\theta}U^{\theta}T^{1-\theta} \leq 2^{\theta}UT^{1-\theta}. \tag{116}$$

The initialization term satisfies

$$(1+U)^2 \leq 4U^2 \leq 4UT^{1-\theta}. \tag{117}$$

Thus, under the range $1 \leq U \leq T^{1-\theta}$,

$$\mathrm{Regret}(K) \leq \widetilde{O}\big(UT^{1-\theta} + TU^{-\gamma}\big). \tag{118}$$

We choose the threshold to balance the two terms as

$$UT^{1-\theta} = TU^{-\gamma}. \tag{119}$$

Rearranging gives $U^{1+\gamma} = T^{\theta}$, so we have

$$U = T^{\frac{\theta}{1+\gamma}} = T^{\frac{\epsilon}{(1+\epsilon)(1+\gamma)}}. \tag{120}$$

This choice lies in the required range because $T \geq 1$, $0 < \theta < 1/2$, and $\gamma > 0$ imply

$$1 \leq T^{\frac{\theta}{1+\gamma}} \leq T^{\theta} \leq T^{1-\theta}. \tag{121}$$

Substituting equation 120 into the first term of equation 118 gives

$$UT^{1-\theta} = T^{1-\theta+\frac{\theta}{1+\gamma}} = T^{1-\frac{\theta\gamma}{1+\gamma}}. \tag{122}$$

The second term has the same power of T.

$$TU^{-\gamma} = T^{1-\frac{\theta\gamma}{1+\gamma}}. \tag{123}$$

Therefore,

$$\mathrm{Regret}(K) \le \widetilde{O}\Big(T^{1-\frac{\theta\gamma}{1+\gamma}}\Big) = \widetilde{O}\Big((KH)^{1-\frac{\epsilon\gamma}{(1+\epsilon)(1+\gamma)}}\Big). \tag{124}$$

For fixed $\epsilon > 0$ and $\gamma > 0$, the exponent is strictly less than one, and the polylogarithmic factors do not change sublinearity. This completes the derivation of the combined regret bound.

G SIMULATION SETTING

In this section, we describe the synthetic heavy-tailed environment, the learning setup, and the evaluation metrics used in Section 8.

Environment. We consider the countably infinite state space $\mathcal{X} = \{0, 1, 2, \ldots\}$ and action set $\mathcal{A} = \{0, 1, 2\}$. State $x = 0$ represents the regular operating state, $x = 1$ represents an automatic inspection state, and states $x \ge 2$ represent positive workload levels. The transition is

$$X' = \begin{cases} 2 + B^{(a)}Z, & x = 0, \\ 2 + Z, & x = 1, \\ [x - d]^+ + C + B^{(q)}Z, & x \ge 2, \end{cases} \tag{125}$$

where $[z]^+ = \max\{z, 0\}$. Conditional on the selected action a,

$$B^{(a)} \sim \mathrm{Bernoulli}(p_a), \qquad (p_0, p_1, p_2) = (10^{-6},\, 2 \times 10^{-6},\, 10^{-4}). \tag{126}$$

For $x \ge 2$, we use

$$C \sim \mathrm{Bernoulli}(s), \qquad B^{(q)} \sim \mathrm{Bernoulli}(q), \qquad s = 0.2, \quad q = 10^{-6}. \tag{127}$$

All external random variables are generated independently at each step.

The burst size Z follows the mixture distribution

$$Z = \begin{cases} M, & \text{with probability } 0.95, \\ \lceil Z^{\mathrm{P}} \rceil, & \text{with probability } 0.05, \end{cases} \tag{128}$$

where $M = 10000$ and Z^{P} follows a Pareto-I distribution with shape parameter $\alpha = 1.5$ and scale 1. Thus, $\Pr(Z^{\mathrm{P}} > z) = z^{-1.5}$ for $z \ge 1$. The Pareto component gives the burst distribution unbounded support and polynomial tail behavior with tail index 1.5. Its mean is $\mathbb{E}[Z] \approx 9500.181$. We set

$$d = M + 2 = 10002. \tag{129}$$

The inspection state $x = 1$ provides repeated observations of large workload states. In particular, large states are rarely reached directly from the regular operating state, but they are also visited through the inspection state. This allows their values to be learned from actual interactions with the environment rather than leaving them as unvisited states with zero-initialized values. All visits to the inspection and workload states, together with their costs and regret, are included in learning.

The cost function is

$$c(x, a) = x + f(a), \tag{130}$$

where

$$\big(f(0), f(1), f(2)\big) = \big(4.2310177,\, 2.2310177,\, 1\big). \tag{131}$$

The cost $f(1)$ is calibrated before the RL evaluation so that, at the reference state $(h, x) = (1, 0)$, action 2 is optimal and action 1 is the next best action with an optimality gap of approximately 0.3. The oracle values at this state are

$$\begin{aligned} Q_1^\star(0, 0) &\approx 7.241, \\ Q_1^\star(0, 1) &\approx 5.250, \\ Q_1^\star(0, 2) &\approx 4.950. \end{aligned} \tag{132}$$

The oracle is used only for this calibration and for evaluation, and is not used by either learning method.

Relation to the heavy-tailed model. The environment has an unbounded state space because of the Pareto component in equation 128. For sufficiently large u, its transition tail is proportional to $u^{-1.5}$. Moreover, for every state and action,

$$|X' - x| \le d + 1 + Z, \tag{133}$$

so the uniform r-th increment moment is finite for any $r \in (1, 1.5)$. The state-dependent cost is $g(x) = x$, which is nonnegative, increasing, and Lipschitz. For sufficiently large x, the expected increment is

$$-d + s + q\mathbb{E}[Z] < 0, \tag{134}$$

which gives negative drift for large workload states. Thus, this synthetic environment is consistent with the heavy-tailed and stability conditions considered in our analysis.

Learning setup. We use horizon $H = 2$ and $K = 10000$ episodes, giving $T = KH = 20000$ learning steps. The initial state is $x_1^1 = 0$, and the physical state continues across episodes as $x_1^{k+1} = x_{H+1}^k$. No state clipping, state aggregation, or episode reset is used.

Both methods initialize all unvisited Q-values to zero and use greedy cost-minimizing action selection, with uniform random selection among exact ties. For a visited tuple (h, x, a), the stored Bellman target is

$$Y = c(x, a) + \min_{a'} Q_{h+1}(X', a'). \tag{135}$$

Once collected, the target is stored with its current value and is not recomputed after later updates.

For each tuple, observations are assigned to $m = 5$ blocks in a round-robin manner. No Q-update is performed until all five blocks are nonempty. Afterward, both methods use

$$Q_h^{\text{new}}(x, a) = (1 - \eta_0) Q_h^{\text{old}}(x, a) + \eta_0 \big(\hat{\mu}_{h,n} - \beta_n\big), \tag{136}$$

with constant learning rate $\eta_0 = 0.45$. For the sample-mean method, $\hat{\mu}_{h,n}$ is the average of all stored targets. For MoM, it is the median of the five block means.

We use $\epsilon = 0.25$, $\delta = (T + 1)^{-4}$, and

$$\beta_n = C_\beta \left(\frac{\log(2/\delta)}{n} \right)^{\epsilon/(1+\epsilon)}. \tag{137}$$

The bonus coefficient for each method is selected using a separate set of 30 tuning runs before the final evaluation. The candidates are

$$\{0.01, 0.05, 0.1, 0.25, 0.5, 1, 2, 3, 5, 7.5, 10, 15, 20, 30, 50\},$$

and the coefficient minimizing the mean cumulative regret over the tuning runs is selected. Both methods select

$$C_\beta = 2. \tag{138}$$

The environment and these coefficients are fixed before the final evaluation. The reported results use 100 new independent runs. Within each run, the two methods use the same underlying external random number streams to provide a paired comparison. The values $m = 5$ and $C_\beta = 2$ are practical simulation parameters. They are chosen for the numerical experiments and are not meant to match the conservative constants required by the theoretical analysis in Theorem 1.

Oracle and cumulative regret. Because $H = 2$, the oracle values can be computed directly. At the last step,

$$V_2^\star(x) = x + \min_a f(a) = x + 1. \tag{139}$$

Thus, at the reference state $x = 0$,

$$Q_1^\star(0, a) = f(a) + 3 + p_a \mathbb{E}[Z]. \tag{140}$$

These values are used only to evaluate the learned policies.

Using the optimal action gap

$$A_h^\star(x, a) = Q_h^\star(x, a) - V_h^\star(x), \tag{141}$$

we define the cumulative regret of run s through episode k as

$$R_s(k) = \sum_{j=1}^{k} \sum_{h=1}^{H} A_h^\star(x_h^{j,s}, a_h^{j,s}). \tag{142}$$

Fig. 2(a) plots

$$\widehat{R}_{\mathrm{avg}}(k) = \frac{1}{N} \sum_{s=1}^{N} \frac{R_s(k)}{kH}, \qquad N = 100. \tag{143}$$

The y-axis is logarithmic. At the end of learning, the mean values are approximately 0.172 for the sample-mean method and 0.116 for MoM, corresponding to a 32.2% reduction in cumulative regret. Across the 100 paired runs, the estimated reduction in cumulative regret remained positive under bootstrap resampling. The corresponding 95% bootstrap interval was approximately $[23.6\%, 39.6\%]$.

Expected policy cost. To measure the quality of the policy learned at a given checkpoint, we freeze the current Q-function and form its greedy policy π. We then compute the expected cumulative cost from the same reference state $x = 0$ and report

$$\frac{V_1^\pi(0)}{H} \tag{144}$$

as the expected cost per step in Fig. 2(b). Because $H = 2$, this expectation can be computed directly from the transition distribution, without additional simulation runs. The optimal reference is

$$\frac{V_1^\star(0)}{H} \approx 2.475. \tag{145}$$

At the final checkpoint, the expected costs per step are approximately 2.550 for the sample-mean method and 2.484 for MoM.

Optimal-action selection. Fig. 2(c) evaluates the action selected by the greedy policy at the reference decision point $(h, x) = (1, 0)$. At each checkpoint, we report the fraction of the 100 runs in which the learned policy selects the true optimal action $a = 2$ at this state. At the final checkpoint, the sample-mean method selects the optimal action in 50 of 100 runs, whereas MoM does so in 94 of 100 runs. This metric refers only to the reference state $(1, 0)$ and should not be interpreted as the optimal-action rate over all states.

All curves in Fig. 2 use all 100 evaluation runs. The shaded regions indicate one standard error around the mean. No smoothing or selection of individual runs is applied.