No-Regret Reinforcement Learning with Heavy-Tailed Rewards

Vincent Zhuang
Tsinghua University
vczhuang@gmail.com

Yanan Sui
Tsinghua University
ysui@tsinghua.edu.cn

Abstract

Reinforcement learning algorithms typically assume rewards to be sampled from light-tailed distributions, such as Gaussian or bounded. However, a wide variety of real-world systems generate rewards that follow heavy-tailed distributions. We consider such scenarios in the setting of undiscounted reinforcement learning. By constructing a lower bound, we show that the difficulty of learning heavy-tailed rewards asymptotically dominates the difficulty of learning transition probabilities. Leveraging techniques from robust mean estimation, we propose Heavy-UCRL2 and Heavy-Q-Learning, and show that they achieve near-optimal regret bounds in this setting. Our algorithms also naturally generalize to deep reinforcement learning applications; we instantiate Heavy-DQN as an example of this. We demonstrate that all of our algorithms outperform baselines on both synthetic MDPs and standard RL benchmarks.

1 Introduction

A wide variety of real-world online decision-making systems generate rewards according to heavy-tailed distributions, such as financial instrument prices, networking routing times, scheduling, hydrology, image and audio noise, localization errors, and others [Georgiou et al., 1999, Hamza and Krim, 2001, Huang and Zhang, 2017, Ruotsalainen et al., 2018]. Since many potential applications of reinforcement learning (RL) are necessarily dependent on such measurements, there is a clear need for RL algorithms that can handle heavy-tailed reward distributions in a provably efficient manner.

Unfortunately, the prevailing assumption in existing reinforcement learning algorithms is that the reward noise is bounded or light-tailed (e.g. sub-Gaussian). Given this, the difficulty of learning the transition probabilities of the Markov decision process (MDP) dominates the difficulty of learning the rewards; in fact, most analyses assume the rewards to be deterministic without loss of generality [Dann and Brunskill, 2015]. As we will demonstrate in this paper, this no longer holds in the heavy-tailed regime.

Although there are no existing RL algorithms that are designed to handle heavy-tailed rewards, several works have studied heavy-tailed feedback in comparatively less general online learning framework. Bubeck et al. [2013] are the first to consider the finite-armed bandit problem in which the reward distributions have only finite $(1 + \epsilon)$-th moments for $\epsilon \in (0, 1]$. They propose the Robust UCB algorithm, which leverages robust mean estimators such as truncated mean and median of means that have tight concentration properties. Under this setting, several extensions of the vanilla bandit problem have been studied: linear bandits [Medina and Yang, 2016, Shao et al., 2018], pure exploration (best-arm identification) [Yu et al., 2018], Lipschitz bandits [Lu et al., 2019], and Bayesian optimization [Chowdhury and Gopalan, 2019]. Thompson sampling has also been analyzed under specific families of heavy-tailed distributions: symmetric α-stable distributions [Dubey and Pentland, 2019] and one-parameter exponential family distributions [Korda et al., 2013].

Many wide-ranging definitions of robustness have been considered in reinforcement learning, including worst-case performance of the learned policy [Smirnova et al., 2019] and adversarial perturbations of the agent [Morimoto and Doya, 2005, Pinto et al., 2017]. The closest works to our setting are those that consider robustness with respect to the reward distributions, e.g. in perturbed or corrupted rewards [Everitt et al., 2017, Wang et al., 2018], or adversarial attacks on the rewards [Weng et al., 2020]. Relatively few works have proposed empirical methods for dealing with noisy reward distributions [Moreno et al., Romoff et al., 2018]. In general,

there is a lack of theoretical and empirical studies on robustness with respect to noisy reward distributions.

The median-of-means estimator is a commonly-used strategy for performing robust mean estimation in heavy-tailed bandit algorithms. In an orthogonal line of work, Pazis et al. [2016] use median-of-means in order to achieve PAC exploration bounds that scale with the variance of the Bellman operator. However, to the best of our knowledge, no prior work has considered our setting of heavy-tailed rewards in the MDP setting.

Our Contribution. We demonstrate that robust mean estimation techniques can be broadly applied to reinforcement learning algorithms (specifically confidence-based methods) in order to provably handle the heavy-tailed reward setting. We instantiate and prove a variety of theoretical guarantees for two representative algorithms, Heavy-UCRL2 and Heavy-Q-Learning. We show via a lower bound construction that in the unbounded heavy-tailed setting, learning the rewards is non-trivial and leads to a regret term that dominates that due to the transition probabilities. We also propose Heavy-DQN to show that our ideas can be extended to the deep reinforcement learning setting. Finally, we empirically demonstrate that our algorithms significantly outperform standard RL algorithms that are not designed to handle heavy-tailed rewards.

2 Problem Statement

We consider a Markov Decision Process (MDP), defined as a tuple $M = (\mathcal{S}, \mathcal{A}, p, R)$ consisting of the state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, and stochastic reward function is $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. An agent begins at some initial state $s_0 \in \mathcal{S}$ (possibly drawn from some distribution), and at each timestep t, selects an action a_t and transitions to state s_{t+1} as dictated by p. The agent also receives a reward $r_t = r(s_t, a_t)$. A policy $\pi : S \to A$ gives an agent's action at any state, and the value function $V^\pi(s)$ denotes the expected return of policy π executed from state s. We also define the Q-value (also state-action value) function $Q^\pi(s, a)$ as the expected return of executing policy π starting with action a at state s.

Typically, r is a stochastic function mapping $\mathcal{S} \times \mathcal{A}$ to a bounded interval, e.g. $[0, 1]$. In this work, we consider MDPs that have reward functions that follow heavy-tailed distributions, and consequently have unbounded range. Following bandit literature [Bubeck et al., 2013], we only assume that the reward distribution has finite $(1 + \epsilon)$-th moments for some $\epsilon \in (0, 1]$. Then the mean reward function is also necessarily finite, which we define as a nonstochastic function $\bar{r} : \mathcal{S} \times \mathcal{A} \to [r_{min}, r_{max}]$. Note that the reward for each state-action pair (s, a) may be drawn from a different distribution, each with some finite $(1 + \epsilon_{s,a})$-th moment. In this case, we can take ϵ to be that of the heaviest distribution, i.e. $\epsilon = \min_{s,a} \epsilon_{s,a}$. The goal of the agent is to maximize the standard undiscounted return: $\sum_t r_t$. An equivalent object is to minimize its *regret*, defined as $\Delta(s_0, T) := \sum_{i=1}^{T} \rho^* - \bar{r}(s_t, a_t)$ where $\rho^* := \max_\pi \lim_{T \to \infty} \mathbb{E}\left[\sum_{t=1}^{T} \bar{r}(s_t, a_t)\right]$.

In order to make the learning problem tractable, we consider two restricted classes of MDPs:

- Communicating MDPs, i.e. MDPs M with finite diameter D defined as $D(M) := \max_{s \neq s'} \min_\pi \mathbb{E}[T(s'|M, \pi, s)]$ where $T(s'|M, \pi, s)$ is the random variable corresponding to the first time step in which state s' is reached under policy π starting from state s.

- Finite-horizon episodic MDPs, in which the agent proceeds in episodes of strictly length H, and the transition probability kernel can change at each step (i.e. we have p_h). The algorithm then runs for time horizon $T = KH$ where K is the number of episodes. In this setting, the diameter is given by H.

3 Algorithms

In this section, we present two algorithms, Heavy-UCRL2 and Heavy-Q-Learning, for reinforcement learning with heavy-tailed rewards. Our algorithms are based on the UCRL2 and Q-learning algorithms respectively, which are representative model-based and model-free algorithms [Jaksch et al., 2010, Jin et al., 2018]). Heavy-UCRL2 and Heavy-Q-Learning leverage ideas from the robust mean estimation literature, which provides estimators with Hoeffding-style concentration bounds under only the assumption of bounded $(1 + \epsilon)$-th moments, for $\epsilon \in (0, 1]$. Confidence-set based algorithms such as UCRL2 and UCFH are immediately amenable to this technique; we also show that the basic idea of truncation also applies to certain bonus-based model-free algorithms such as Q-learning. We also propose a Heavy-DQN, a direct extension of Heavy-Q-Learning to the deep RL setting.

We first briefly review UCRL2 and Q-learning. UCRL2 is a model-based RL algorithm that implements the broad principle of optimism in the face of uncertainty. At the beginning of each episode, UCRL2 constructs confidence sets for all transition probabilities and rewards (eqs. 1 and 2), which comprise an (infinite) set of statistically plausible MDPs $\mathcal{M}$. UCRL2 then compute an optimistic MDP $\widetilde{M} \in \mathcal{M}$ via extended value iteration, which maximizes the optimal average reward.

It then executes the optimal policy on $\widetilde{M}$, $\widetilde{\pi}$, for the entirety of the next episode.

$$\|(\widetilde{r} - \hat{r}_k)(s,a)\| \leq \sqrt{\frac{7\log\left(\frac{2SAt_k}{\delta}\right)}{2\hat{N}_k(s,a)}} \qquad (1)$$

$$\|\widetilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \leq \sqrt{\frac{14S\log\left(\frac{2At_k}{\delta}\right)}{\hat{N}_k(s,a)}} \qquad (2)$$

Notation. We use the subscript k to denote the k-th episode. Hence, t_k denote the total number of iterations at the start of episode k, $v_k(s,a)$ denotes the state-action visitation count within episode k, and $N_k(s,a)$ denotes the state-action visitation up to episode k. We use $\hat{r}(s,a)$ and $\bar{r}(s,a)$ to denote the algorithm's estimated mean for the reward and the actual mean reward respectively.

Q-learning. Q-learning directly learns the Q-values by repeatedly applying the following one-step Bellman update:

$$Q_h(s,a) = (1-\alpha_t)Q_h(s,a) + \alpha_t[r_h(s,a) + V_{h+1}(s') + b_t] \qquad (3)$$

where Q_h and V_h are the state and state-action value functions at step $h \leq H$, t is the total number of times the transition (s,a) has been taken thus far, $\alpha_t \in (0,1]$ is the learning rate, and b_t is an exploration bonus. Jin et al. [2018] show that Q-learning has near-optimal regret bounds in the episodic MDP setting if the Q-values are optimistically initialized to H and α_t and b_t are chosen appropriately. Specifically, they use learning rate $\alpha_t := \frac{H+1}{H+t}$, and propose two possible exploration bonuses. The first, UCB-Hoeffding, follows from an application of the Azuma-Hoeffding inequality: $c\sqrt{\frac{H^3\iota}{t}}$ where $\iota := \log(SAT/\delta)$. The second, UCB-Bernstein leverages the Freedman inequality, which allows for a tighter confidence interval based on the empirical estimate of the variance of the next-state value.

3.1 Heavy-tailed Mean Estimation

Robust mean estimators. We define a robust mean estimator to be an estimator $\hat{\mu}$ that satisfies the following property. Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite mean μ, $\epsilon \in (0,1]$ and v a bound on the centered $(1+\epsilon)$-th moment of X, i.e. $\mathbb{E}[(X-\mu)^{1+\epsilon}] \leq v$ (we will also denote an upper bound on the uncentered $(1+\epsilon)$-th moment of X by $u \geq \mathbb{E}[X^{1+\epsilon}]$). Then, for all $\delta \in (0,1)$ there exists a constant c such that with probability at least $1-\delta$,

$$|\mu - \hat{\mu}| \leq v^{\frac{1}{1+\epsilon}}\left(\frac{c\log(1/\delta)}{n}\right)^{\frac{\epsilon}{1+\epsilon}} \qquad (4)$$

We note that if $\epsilon = 1$, i.e. X has finite variance, then v is simply the variance factor, and we recover a Hoeffding-style bound on the deviation of $\hat{\mu}$. Further details can be found in Bubeck et al. [2013]. We now review two major robust mean estimators that satisfy eq. 4:

- **Truncated empirical mean.** This estimator truncates samples with too high magnitude to zero:

$$\hat{\mu}_T = \frac{1}{n}\sum_{t=1}^{n} X_t \mathbb{1}_{\left\{|X_t| \leq \left(\frac{ut}{\log(\delta^{-1})}\right)^{\frac{1}{1+\epsilon}}\right\}} \qquad (5)$$

  where $c = 4^{(1+\epsilon)/\epsilon}$.

- **Median-of-means.** This simple estimator simply splits the n samples into k blocks and takes the median of the mean in each block:

$$\hat{\mu}_M = \text{median}\left(\left\{\frac{1}{N}\sum_{t=(i-1)N+1}^{iN} X_t\right\}_{i \leq N}\right) \qquad (6)$$

  where $N = \lceil\frac{n}{k}\rceil$, $k = \lfloor 8\log(e^{1/8}/\delta) \wedge n/2\rfloor$ and $c = 32\sqrt[6]{12}$.

Failure of empirical mean/UCRL2 for heavy-tailed rewards. We first demonstrate that naively applying vanilla UCRL2 (i.e. with the empirical mean estimator for the rewards) to the heavy-tailed setting results in either a statistically-incorrect or vacuous algorithm. As shown by Bubeck et al. [2013], the tightest possible confidence interval for the empirical mean scales as $O\left(\left(\frac{1}{\delta t^\epsilon}\right)^{\frac{1}{1+\epsilon}}\right)$. Since the analysis of UCRL2 requires that δ inversely scales polynomially in t, the appropriate confidence interval for the rewards scales polynomially in t, which clearly dominates the (constant) reward means. Thus, attempting to construct a version of empirical-mean UCRL2 with guarantees in the heavy-tailed setting results in an algorithm that eventually devolves into random exploration.

3.2 Heavy-UCRL2

We propose Heavy-UCRL2, which utilizes a robust estimator for $r(s,a)$ instead of the empirical mean. This allows us to use the following confidence intervals for the rewards:

$$\|\widetilde{r}(s,a) - \hat{r}_k(s,a)\| \leq v^{\frac{1}{1+\epsilon}}\left(\frac{7c\log(2SAt_k/\delta)}{\max\{1, N_k(s,a)\}}\right)^{\frac{\epsilon}{1+\epsilon}}, \qquad (7)$$

where c is given as in eq. 25. The confidence intervals for the transition probabilities remain unchanged from eq. 2. The pseudocode for Heavy-UCRL2 is given in Algorithm 1.

Algorithm 1 Heavy-UCRL2

1: input: confidence parameter $\delta \in (0,1)$, $\mathcal{S}$, $\mathcal{A}$, $\hat{r}$ a robust mean estimator
2: $t \leftarrow 1$, initial state s_1
3: **for** episodes $k = 1, 2, \dots$ **do**
4:    $t_k \leftarrow t$, set $N_k(s,a)$
5:    For all $(s,a) \in S \times A$ initialize $v_k(s,a)$ to 0
6:    Set $\mathcal{M}_k$ to be the set of all MDPs with states S and actions A with transitions close to $\hat{p}_k(\cdot|s,a)$ and rewards close to $\hat{r}(s,a)$ according to eqs. 2 and 7
7:    Compute policy $\widetilde{\pi}_k$ using extended value iteration on optimistic MDP $\tilde{M}_k = (\mathcal{S}, \mathcal{A}, \tilde{P}_k, \tilde{r}_k)$.
8:    Execute $\widetilde{\pi}_k$ for episode k until $v_l(s_t, \widetilde{\pi}_k(s_t)) = \max\{1, N_k(s_t, \widetilde{\pi}_k(s_t))\}$
9: **end for**

Algorithm 2 Heavy-Q-Learning

1: input: confidence parameter $\delta \in (0,1)$, $\mathcal{S}$, $\mathcal{A}$, bonus b_t either UCB-Hoeffding or UCB-Bernstein, r_{max} the maximum possible mean reward
2: initialize $Q_h(s,a) \leftarrow H r_{max}$ and $N_h(s,a) \leftarrow 0$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$
3: **for** episode $k = 1, \dots, K$ **do**
4:    **for** episodes $h = 1, \dots, H$ **do**
5:       Take action $a_h \leftarrow \arg\max_{a'} Q_h(s_h, a')$
6:       $t = N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$; $b'_t \leftarrow b_t + c_2 H u^{\frac{1}{1+\epsilon}} \left(\frac{\iota}{t}\right)^{\frac{\epsilon}{1+\epsilon}}$
7:       $Q_h(s_h, a_h) \leftarrow (1 - \alpha_t) Q_h(s_h, a_h) + \alpha_t[r_h(s_h, a_h)\mathbb{1}_{|r_h(s,a)| \leq B_t} + V_{h+1} + b'_t]$
8:    **end for**
9: **end for**

Comparison with UCRL2. Here, we briefly emphasize the value of having tight statistically-valid confidence intervals. For UCRL2 with the empirical mean estimator for the reward, there are two possibilities for the confidence interval. One can either use the tight confidence interval for the bounded/sub-Gaussian noise case, or a statistically-valid yet loose bound from Bubeck et al. [2013]. In the latter case, the $1/\delta$ factor is not hidden inside a logarithm and hence causes the confidence intervals to blow up polynomially in t. In the former case, we empirically demonstrate in section 5 that under-estimating the confidence interval can degrade the performance of the algorithm. In general, it is clear that using statistically incorrect confidence intervals can cause the algorithm to fail, and having tight confidence intervals is always better, as to avoid unnecessary exploration.

3.3 Heavy-Q-Learning

Our Heavy-Q-Learning algorithm, detailed in Algo-

rithm 2, extends the basic Q-learning algorithm in a simple way: instead of using the reward directly in the Bellman update, we truncate it to zero if the reward has magnitude too large, and augment the exploration bonus with a term designed to deal with the heavy-tailedness of the rewards. Hence, the Heavy-Q-Learning update can be written as follows:

$$Q_h(s,a) = (1-\alpha_t)Q_h(s,a) + \alpha_t[r_h(s,a)' + V_{h+1}(s') + b'_t] \quad (8)$$

where

$$r_h(s,a)' := r_h(x,a)\mathbb{1}_{|r_h(s,a)| \leq B_t} \quad (9)$$

and

$$b'_t := b_t + 8H u^{\frac{1}{1+\epsilon}} \left(\frac{\log(2SAT/\delta)}{t}\right)^{\frac{\epsilon}{1+\epsilon}}. \quad (10)$$

We define $B_t := \left(\frac{ut}{\log(2SAT/\delta)}\right)^{\frac{1}{1+\epsilon}}$ as in the truncated mean estimator (eq. 5); we use $\delta' = \frac{\delta}{2SAT}$ in order to facilitate a union bound over S, A, T. The base exploration bonus b_t is either of the original UCB-Hoeffding or UCB-Bernstein bonuses proposed in Jin et al. [2018] scaled by r_{max}. Our algorithm is compatible with both bonuses since they are primarily designed to deal with the uncertainty in transition probabilities.

As we will show in the following section, this simple modification allows Heavy-Q-Learning to be robust to heavy-tailed rewards while also preserving all of the advantages of standard Q-learning over UCRL2, such as its smaller time and space complexities. We note that for this reason only the truncated mean estimator is amenable to Q-learning; the median-of-means estimator would require recomputing the entire history at each step. Finally, an additional modification we make in Heavy-Q-Learning is to optimistically initialize the Q-function to uniformly be $H r_{max}$. This is one disadvantage of Heavy-Q-Learning over Heavy-UCRL2 (and in general Q-Learning vs UCRL2): the former requires explicit a priori knowledge of r_{max}, whereas the latter does not.

3.4 Heavy-DQN

Recently, there has been much focus on deep reinforcement learning algorithms and applications. To demonstrate that the general ideas behind robust mean estimation can be extended to this setting, we propose Heavy-DQN as a combination of Heavy-Q-Learning and DQN for discrete-action deep reinforcement learning [Mnih et al., 2013]. To perform adaptive truncation in non-tabular environments, we use methods from the count-based exploration literature, e.g. the SimHash function ϕ [Tang et al., 2017]. Since we want to estimate the count per state-action pair, we consider the

SimHash counts per action. Then, our reward truncation threshold is given by $(C \cdot n(\phi(s,a)))^{\frac{1}{1+\epsilon}}$, where C is an environment-dependent hyperparameter. Although there is recent work on extending optimistic initialization and UCB exploration from Q-learning to the deep RL setting, in our experiments it was sufficient to use standard Q-learning initialization and linearly-decaying epsilon-greedy exploration [Rashid et al., 2020].

Connections to reward clipping. Deep RL algorithm implementations commonly clip rewards to some fixed range $[-x, x]$ or even as $\text{sign}(r)$ [Mnih et al., 2013]. The reward truncation in Heavy-DQN can be viewed as an adaptive version of this kind of fixed clipping. We note the differing original intents of these mechanisms: the main purpose of reward clipping is to stabilize the training dynamics of the neural networks, whereas our method is designed to ensure theoretically-tight reward estimation in the heavy-tailed setting for each state-action pair. Whether there are any unifying connections between these two purposes is an open area for future work.

4 Theoretical Results

In this section, we present a variety of theoretical results for Heavy-UCRL2 and Heavy-Q-Learning as well as a general lower bound for the heavy-tailed RL setting. In particular, for Heavy-UCRL2 and Heavy-Q-Learning with both bonuses, we prove standard minimax upper bounds on the expected regret. In addition, for Heavy-UCRL2 we prove analogous results for every result in Jaksch et al. [2010]: a regret bound dependent on the gap between the best and second-best policies and a regret bound for a MDP that is allowed to change a fixed number of times. As noted in [Jin et al., 2018], our regret bounds also naturally translate to PAC bounds.

In all the following, we denote $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$, and use ι to hide the logarithmic factor $\log\left(\frac{2SAT}{\delta}\right)$. For the sake of conciseness, we primarily state the theorems in this section, and defer the detailed proofs to the appendix.

4.1 Heavy-UCRL2 results

Minimax regret bound. We show that Heavy-UCRL2 enjoys an upper bound on worst-case regret of $\tilde{O}(DS\sqrt{AT} + (SAT)^{\frac{1}{1+\epsilon}})$ for sufficiently large T. Note that if the variance is bounded, we recover an identical bound as in the bounded rewards case: $\tilde{O}(DS\sqrt{AT})$. This is formalized in the following theorem:

Theorem 1. *Let $R_\Delta := r_{max} - r_{min}$. With probability*

at least $1 - \delta$, the regret of Heavy-UCRL2 is bounded by

$$20R_\Delta DS\sqrt{AT \log\left(\frac{T}{\delta}\right)} + \tag{11}$$
$$(2C_\epsilon + 1)v^{\frac{1}{1+\epsilon}}(7c\iota)^{\frac{\epsilon}{1+\epsilon}}(SAT)^{\frac{1}{1+\epsilon}}$$

The first term corresponds to the regret incurred due to the uncertainty in transition probabilities, and the second term to the regret due to the heavy-tailed nature of the rewards. In the standard bounded reward setting, the first term is asymptotically dominant. However, since D, S, A are fixed for any given MDP, it is clear that the heavy-tailed regret term will dominate for sufficiently large T, which matches the minimax regret order of T for linear bandits [Shao et al., 2018].

We also note that there is an unavoidable $R_\Delta := r_{max} - r_{min}$ multiplicative factor, where $r_{max} = \max_{s,a} r(s,a)$ and $r_{min} = \min_{s,a} r(s,a)$. However, such a factor would also be present in the trivial generalization from bounded $[0,1]$ reward distributions to $[r_{min}, r_{max}]$.

Theorem 1 also immediately gives the following PAC bound on the regret:

Corollary 1. *With probability at least $1 - \delta$, the average per-step regret of Heavy-UCRL2 is at most λ for any*

$$T \geq \max\left(4^2 \cdot 20^2 \frac{R_\Delta^2 D^2 S^2 A}{\lambda^2} \log\left(\frac{40 R_\Delta DSA}{\delta\lambda}\right),\right.$$
$$\left. \alpha \log\left(\frac{2SA}{\delta}\right) + 2\alpha \log\left(\frac{\alpha}{\delta}\right)\right) \tag{12}$$

where

$$\alpha := \frac{7c(4C_\epsilon + 2)^{\frac{1+\epsilon}{\epsilon}} v^{\frac{1}{\epsilon}}(SA)^{\frac{1}{\epsilon}}}{\lambda^{\frac{1+\epsilon}{\epsilon}}}.$$

Problem-dependent regret bound. We also have a logarithmic regret bound dependent on the gap between the best and second-best policies, as in UCRL2. To simplify the statement of the theorem, we only consider the heavy-tailed regime, i.e. that in which equation 4 is satisfied. We use λ to denote gaps in reward values, to differentiate from ϵ for the reward distribution.

Theorem 2. *For any initial state $s \in \mathcal{S}$, any T satisfying equation 4, and any $\lambda > 0$, with probability at least $1 - 3\delta$ the regret of Heavy-UCRL2 is*

$$\Delta(M, s, T) \leq 7c\iota (4C_\epsilon + 2)^{\frac{1+\epsilon}{\epsilon}} \left(\frac{SA}{\lambda}\right)^{1/\epsilon} + \lambda T$$

Furthermore, let $g := \rho^*(M) - \max_{s \in S} \max_{\pi:\mathcal{S}\to\mathcal{A}} \{\rho(M, \pi, s) : \rho(M\pi, s) > \rho^*(M)\}$

be the gap between the average reward of the best and second-best policies. Then, the expected regret of Heavy-UCRL2 (with parameter $\delta := \frac{1}{3T}$) is bounded by

$$\mathbb{E}[\Delta(M,s,T)] \leq 7c\iota \left(4C_\epsilon + 2\right)^{\frac{1+\epsilon}{\epsilon}} \left(\frac{2SA}{g}\right)^{1/\epsilon} +$$

$$\sum_{s,a} \left\lceil 1 + \log_2 \left(\max_{\pi:\pi(s)=a} T_\pi\right)\right\rceil \max_{\pi:\pi(s)=a} T_\pi$$

Theorem 2 highlights that as ϵ becomes arbitrarily small, the problem-dependent regret bound goes to infinity. In theorem 6, we show that this dependency is inevitable.

Regret under changing MDP. Heavy-UCRL2 is designed to be robust to heavy-tailed reward distributions. In this section, we consider an orthogonal notion of robustness. Namely, if the MDP is allowed to change up to $\ell - 1$ times after the initial MDP, under the condition that the diameter never exceeds D, we have the following upper bound on the regret of Heavy-UCRL2.

Theorem 3. *Suppose $\epsilon < 1$ and $T \in O\left(\frac{D^{\frac{2+2\epsilon}{1-\epsilon}} S^{\frac{2\epsilon}{1-\epsilon}}}{A}\right)$ Restarting Heavy-UCRL2 with parameter $\frac{\delta}{\ell^2}$ at steps $\left\lceil \frac{i^{(1+2\epsilon)/\epsilon}}{\ell^{(1+\epsilon)/\epsilon}}\right\rceil$ for $i = 1, 2, 3, \ldots$, the regret of Heavy-UCRL2 is upper bounded by*

$$R_\Delta \ell^{\frac{\epsilon}{1+2\epsilon}} T^{\frac{1+\epsilon}{1+2\epsilon}} (SA)^{\frac{1}{1+\epsilon}} \tag{13}$$

with probability at least $1 - \delta$.

We note that this theorem skips the case where $\epsilon = 1$, which recovers the result of Theorem 6 in Jaksch et al. [2010] up to a constant multiplicative factor.

4.2 Heavy-Q-Learning Results

We first state the following regret bounds for Heavy-Q-Learning with UCB-Hoeffding and UCB-Bernstein:

Theorem 4. *In the finite-horizon episodic MDP setting, the regret of Heavy-Q-Learning with UCB-Hoeffding is $\Theta\left(r_{max} H^2\sqrt{SAT} + H^2(SA\iota)^{\frac{1}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}\right)$.*

Theorem 5. *In the finite-horizon episodic MDP setting, the regret of Heavy-Q-Learning with UCB-Bernstein is*

$$\Theta\Big(\sqrt{H^3 r_{max}^3 SAT\iota} + H^2(SA\iota)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$$
$$+\sqrt{H^9 r_{max}^2 u^{\frac{1}{1+\epsilon}} S^3 A^3 \iota^3} + \sqrt{H^{\frac{1+4\epsilon}{\epsilon}} r_{max} S^2 A^2 \iota^2}$$
$$+ H^{\frac{1+3\epsilon}{\epsilon}}\sqrt{S^3 A^3 \iota^4 \epsilon}\Big).$$

Again, we note that that for sufficiently large T, the heavy-tailed regret term dominates; i.e. learning the
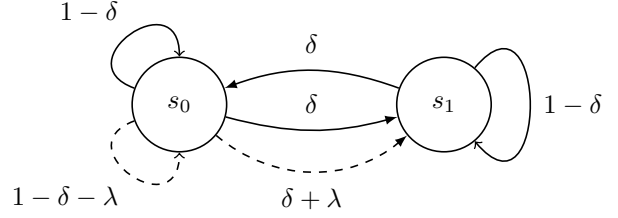

Figure 1: Two-state MDP for lower bound.

rewards is more difficult than learning the transition probabilities. As expected, UCB-Bernstein exploration offers no improvement on the asymptotic regret.

Notably, aside from the simple scaling due to r_{max}, our algorithm only necessitates another additive factor corresponding to the second term in the regret bounds (UCB-Bernstein also has some terms independent of T). Interestingly, our analysis shows that even though Q-learning does not explicitly perform mean estimation and is not amenable to all robust mean estimation techniques, the truncation technique can still be applied. This indicates that truncation can potentially be used to make a broad class of optimistic confidence-interval based algorithms robust in the heavy-tailed setting.

Furthermore, although this setting is slightly different from that of Heavy-UCRL2, it is clear that the dependence on SA is better for Heavy-Q-Learning by a factor of $(SA)^{\frac{1-\epsilon}{1+\epsilon}}$. As we will show, the dependence of this bound on S, A, T is tight. Finally, as for Heavy-UCRL2, this minimax regret bound also naturally translates into a PAC bound.

4.3 Lower bound

We prove the following lower bound on the expected regret for any algorithm in this heavy-tailed MDP setting.

Theorem 6. *For any fixed T and algorithm, there exists a communicating MDP M with diameter D such that the expected regret of the algorithm is $\Omega\left((SA)^{\frac{\epsilon}{1+\epsilon}}(T)^{\frac{1}{1+\epsilon}}\right)$. In the finite-horizon episodic setting, there exists a MDP such that the expected regret is $\Omega\left(H(SA)^{\frac{\epsilon}{1+\epsilon}}(T)^{\frac{1}{1+\epsilon}}\right)$.*

Intuitively, the proof of this theorem follows by combining the lower bound constructions for (problem-independent) heavy-tailed bandits in Bubeck et al. [2013] and finite-diameter MDPs in Jaksch et al. [2010]. The latter considers, without loss of generality, a simple two-state MDP (shown in figure 1) in which for all actions except a single optimal action a^*, the probability of transitioning to the other state is δ. However, when
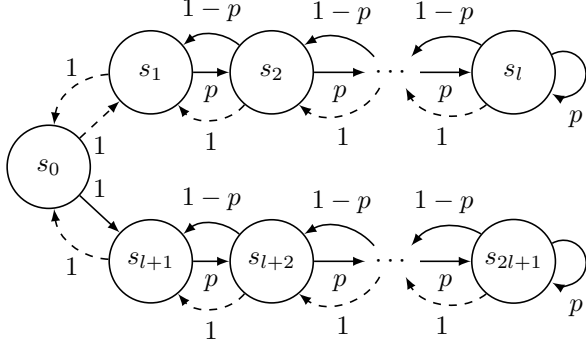
Figure 2: DoubleChain MDP. The dashed and solid lines indicate the transition probabilities of the two actions.

action a^* is taken in s_0, the agent has probability $\delta + \lambda$ of transitioning to s_1.

Now, recall that in the construction in Bubeck et al. [2013], there are $K - 1$ arms have Bernoulli reward scaled by $\frac{1}{\gamma}$ with parameter $\gamma^{1+\epsilon} - \Delta\gamma$ and one optimal arm with $\frac{1}{\gamma}$-scaled Bernoulli with parameter $\gamma^{1+\epsilon}$, where Δ is the gap in the expected reward of the optimal arm and that of other arms, and $\gamma := (2\Delta)^{1/\epsilon}$. Then, directly setting $\Delta = (K/T)^{\frac{\epsilon}{1+\epsilon}}$ yields a $K^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$ regret bound. In our setting, noting that a two-state MDP with unknown transition probabilities is equivalent to learning a Bernoulli reward [Osband and Van Roy, 2016], translating the bandit lower-bound construction to the MDP in figure 1 yields a lower bound of $\Theta\left((SA)^{\frac{\epsilon}{1+\epsilon}}(T)^{\frac{1}{1+\epsilon}}\right)$.

For the episodic setting, we can concatenate H of these MDPs together into a chain, as in Jin et al. [2018]. Hence, the total number of samples available for each step $h \leq H$ is now T/H, and we can rescale the bound accordingly.

5 Experimental Results

5.1 Tabular MDPs

We first evaluate Heavy-UCRL2 and Heavy-Q-Learning on synthetic MDPs (SixArms [Strehl and Littman, 2008] and DoubleChain, shown in figure 2), comparing against classical light-tailed baselines: Gaussian Posterior Sampling (PSRL), Q-learning, and UCRL2. For PSRL, it is non-trivial to design a variant that accomodates general heavy-tailed reward distributions. However, on light-tailed distributions, it is state-of-the-art [Osband and Van Roy, 2017]. Hence, we include it as a benchmark in order to explicitly demonstrate that an algorithm designed for light-tailed rewards may perform poorly in the heavy-tailed setting. For UCRL2,

to ensure a fair comparison, we use the heavy-tailed variant described in section 3. As discussed, this algorithm quickly devolves into random exploration, i.e. uniform selection of actions, and hence performs extremely poorly in our experiments. We therefore omit heavy-tailed UCRL2 from our following results. For all of the above algorithms, we scale the confidence intervals by factors searched within $[1e - 2, 1]$, as is standard practice [Lu et al., 2019].

For Heavy-UCRL2 and Heavy-Q-Learning, we used the truncated mean estimator, and $\epsilon = 0.05$ in all experiments. For Heavy-Q-Learning, we use the UCB-Hoeffding bonus.

For the reward distributions, we used symmetric Levy α-stable distributions, for which the heaviness of the tail is controlled by α and the mean can be arbitrarily specified. In particular, the $(1 + \epsilon)$-th moments of α-stable distributions for $\epsilon < \alpha$ are bounded, i.e. these distributions are the "heaviest" possible in this setting [Dubey and Pentland, 2019]. Such distributions can be denoted as $\mathcal{L}(\mu, \alpha, \beta, \sigma)$, where μ is the mean, and α, β, σ the stable, skew, and shape parameters respectively. In all experiments, we set $\alpha = 1.1$, and only consider such distributions with $\sigma = 1$ that are symmetric (i.e. $\beta = 0$). The full experimental details, including the specific reward distributions for each state-action pair, can be found in the appendix.

Results. We report the total cumulative rewards for each enviroment, averaged over 30 random seeds, in figures 3a and 3b. Heavy-UCRL2 and Heavy-Q-Learning clearly outperform PSRL and Q-learning on both MDPs. As expected, the latter two converge too quickly to a suboptimal policy, and never succeed in identifying the best transition. On the other hand, Heavy-UCRL2 and Heavy-Q-Learning are able to perform enough exploration to discover better policies due to their appropriately-constructed confidence regions/bonuses. This is evidenced by the early iterations, in which PSRL and Q-learning initially achieve higher reward but ultimately fail to identify the best policies.

Overall, these experiments underscore two complementary failures of light-tailed algorithms in heavy-tailed settings: they may converge too quickly to a suboptimal policies, and thus will fail to allocate enough samples to the optimal, heavy-tailed arms because of 1. the brittleness of the empirical mean estimator and 2. the too-tight confidence intervals.

5.2 Deep Reinforcement Learning Experiments

We tested our deep RL algorithm Heavy-DQN on the classical control environment of MountainCar. We

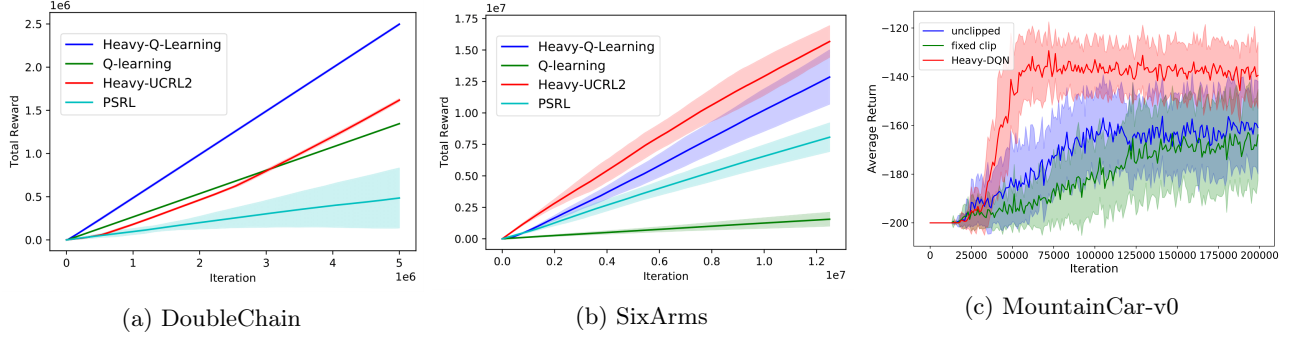(a) DoubleChain   (b) SixArms   (c) MountainCar-v0

Figure 3: Mean cumulative returns averaged over independent trials, with ± 1 standard deviation shaded.

perturbed the standard rewards with heavy-tailed noise following a symmetric α-stable distribution with $\alpha = 1.1$ and scale parameter 5. The algorithm conservatively uses $\epsilon = 0.05$. Our code is based off of the Dopamine reinforcement learning library [Castro et al., 2018], and the remaining hyperparameters were untuned from the defaults (restated in the appendix).

We compare our adaptive truncation method vs. a naive fixed clipping baseline, as well as an unclipped baseline, where the only hyperparameters tuned are the truncation coefficient c as well as the clipping range. All other shared hyperparameters are held constant between algorithms. To ensure fair comparison, each algorithm is run over a fixed set of 24 random seeds.

Results. We plot the mean training return with ± 1 standard deviations shaded in figures 3c. The best c was 0.015, and the best fixed clipping range was $[-5, +5]$. As expected, our adaptive truncation method outperforms both the fixed clipping and unclipped baseline. Heavy-DQN is able to quickly converge to a near-optimal policy, whereas the others become stuck in sub-optimal ones. Overall, these experiments demonstrate that truncation has promise in robustifying deep reinforcement learning algorithms with respect to reward noise.

6   Conclusion & Discussion

We introduced the heavy-tailed reinforcement learning problem and proposed two algorithms, Heavy-UCRL2 and Heavy-Q-Learning, that are provably efficient in this setting. More importantly, our results indicate that a broad-class of confidence-based RL algorithms are amenable to simple techniques from robust statistics. For example, state-of-the-art algorithms such as UCFH and UCBVI are likely compatible with the ideas behind Heavy-UCRL2 and Heavy-Q-Learning [Dann and Brunskill, 2015, Azar et al., 2017]. However, it is unclear how such ideas can be applied to Bayesian

algorithms, such as PSRL [Agrawal and Jia, 2017].

All of our algorithms require a priori knowledge of ϵ. However, this is a drawback of all bandit algorithms with heavy-tailed noise, and probably cannot be avoided in the absence of further assumptions [Shao et al., 2018]. We leveraged the median-of-means and truncated mean estimators for their theoretical guarantees, but both have their drawbacks in practical usage. The truncated mean estimator is simple to use and requires no overhead, but the confidence interval scaling factor can be large if the magnitude of the mean rewards are large, due to its dependence on the raw $(1 + \epsilon)$-th moment. The median-of-means estimator is typically more robust and uses the centered moment bound, but has suboptimal constants. It also requires significantly larger time/space complexity (an additional multiplicative $O(T)$ factor for both).

Overall, we showed that our algorithms bridged theory and practice, providing significant empirical evidence that Heavy-UCRL2 and Heavy-Q-Learning outperform classical RL algorithms on MDPs with heavy-tailed rewards even without relaxing their theoretical assumptions. We also proposed Heavy-DQN, an extension of Heavy-Q-Learning to the deep reinforcement learning setting. Our results demonstrate that despite its simplicity, this adaptive clipping method is effective in stabilizing DQN under heavy-tailed reward noise. There are a number of straightforward extensions that may improve Heavy-DQN's performance and help it generalize to higher-dimensional environments, such as learned hash codes [Tang et al., 2017].

In general, we proposed noisy rewards as a problem setting with real-world relevance. This, and its connections to clipping methods that are prevalent in RL application implementations, have been under-explored thus far and merit further study.

References

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

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

Pablo Samuel Castro, Subhodeep Moitra, Carles Gelada, Saurabh Kumar, and Marc G. Bellemare. Dopamine: A Research Framework for Deep Reinforcement Learning. 2018. URL http://arxiv.org/abs/1812.06110.

Sayak Ray Chowdhury and Aditya Gopalan. Bayesian optimization under heavy-tailed payoffs. In *Advances in Neural Information Processing Systems*, 2019.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

Maria Dimakopoulou and Benjamin Van Roy. Coordinated exploration in concurrent reinforcement learning. *arXiv preprint arXiv:1802.01282*, 2018.

Abhimanyu Dubey and Alex Pentland. Thompson sampling on symmetric α-stable bandits. *arXiv preprint arXiv:1907.03821*, 2019.

Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. Reinforcement learning with a corrupted reward channel. *arXiv preprint arXiv:1705.08417*, 2017.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Panayiotis G. Georgiou, Panagiotis Tsakalides, and Chris Kyriakakis. Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise. *IEEE transactions on multimedia*, 1(3):291–301, 1999.

A Ben Hamza and Hamid Krim. Image denoising: A nonlinear robust statistical approach. *IEEE transactions on signal processing*, 49(12):3045–3054, 2001.

Yulong Huang and Yonggang Zhang. A new process uncertainty robust student'st based kalman filter for sins/gps integration. *IEEE Access*, 5:14391–14404, 2017.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in neural information processing systems*, pages 1448–1456, 2013.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning*, pages 4154–4163, 2019.

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

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Alvaro Moreno, José D Martín, Emilio Soria, and Rafael Magdalena. Noisy reinforcements in reinforcement learning: some case studies based on gridworlds.

Jun Morimoto and Kenji Doya. Robust reinforcement learning. *Neural computation*, 17(2):335–359, 2005.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2701–2710. JMLR. org, 2017.

Jason Pazis, Ronald E Parr, and Jonathan P How. Improving pac exploration using the median of means. In *Advances in Neural Information Processing Systems*, pages 3898–3906, 2016.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2817–2826. JMLR. org, 2017.

Tabish Rashid, Bei Peng, Wendelin Böhmer, and Shimon Whiteson. Optimistic exploration even with a pessimistic initialisation. *arXiv preprint arXiv:2002.12174*, 2020.

Joshua Romoff, Peter Henderson, Alexandre Piché, Vincent Francois-Lavet, and Joelle Pineau. Reward

estimation for variance reduction in deep reinforcement learning. *arXiv preprint arXiv:1805.03359*, 2018.

Laura Ruotsalainen, Martti Kirkko-Jaakkola, Jesperi Rantanen, and Maija Mäkelä. Error modelling for multi-sensor measurements in infrastructure-free indoor navigation. *Sensors*, 18(2):590, 2018.

Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems*, pages 8420–8429, 2018.

Elena Smirnova, Elvis Dohmatob, and Jérémie Mary. Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*, 2019.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in neural information processing systems*, pages 2753–2762, 2017.

Jingkang Wang, Yang Liu, and Bo Li. Reinforcement learning with perturbed rewards. *arXiv preprint arXiv:1810.01032*, 2018.

Tsui-Wei Weng, Krishnamurthy (Dj) Dvijotham*, Jonathan Uesato*, Kai Xiao*, Sven Gowal*, Robert Stanforth*, and Pushmeet Kohli. Toward evaluating robustness of deep reinforcement learning with continuous control. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylL0krYPS.

Xiaotian Yu, Han Shao, Michael R Lyu, and Irwin King. Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *UAI*, pages 937–946, 2018.

A   Proof of Theorem 1

A.1   Preliminaries

Overview of Heavy-UCRL2 proof strategy.   Our regret analysis for Heavy-UCRL2 follows the standard strategy for optimistic algorithms [Osband and Van Roy, 2017, Jaksch et al., 2010]. We first bound the regret in terms of the regret incurred in each episode and the random deviation of the rewards. We then consider the regret from episodes in which the confidence regions do not hold, i.e. the true MDP lies outside the confidence region. Finally, we bound the regret when the confidence regions do hold. This regret can be decomposed into the sum of three terms: the regret due to the error in transition probabilities, regret due to the error in rewards, and the regret due to extended value iteration.

The main differences in the analysis are due to the unboundedness and heavy-tailed nature of the rewards. From section 3, we show that for reward distributions with bounded $(1 + \epsilon)$-th moments ($\epsilon \in (0, 1]$), we have concentration bounds similar to that of Hoeffding's inequality for arbitrary distributions over $[0, 1]$. We also leverage the boundedness of the 1st moment (i.e. the mean) to bound the regret when the confidence intervals fail. There is also an unavoidable $R_\Delta := r_{max} - r_{min}$ multiplicative factor, where $r_{max} = \max_{s,a} r(s, a)$ and $r_{min} = \min_{s,a} r(s, a)$. Note that although the observed rewards r_t are unbounded, the means $\bar{r}$ necessarily are, since by assumption the $(1 + \epsilon)$-th moment and hence the mean are bounded.

Convergence of extended value iteration for heavy-tailed rewards.   Since Heavy-UCRL2 relies on extended value iteration, we need to show that it works in our heavy-tailed setting. Fortunately, since extended value iteration uses fixed (optimistic) reward values, Theorem 7 in [Jaksch et al., 2010] holds identically. We restate it below for completeness:

Theorem 7. *Let $\mathcal{M}$ be the set of MDPs with state space $\mathcal{S}$, action space $\mathcal{A}$, transition probabilities $\tilde{p}(\cdot|s, a)$ and mean rewards $\tilde{r}(s, a)$ that satisfy eqs. 2 and 7, and $u_i(s)$ denote the state value of s at iteration i. Then, if $\mathcal{M}$ contains at least one communicating MDP, extended value iteration converges. Furthermore, stopping extended value iteration when*

$$\max_{s \in \mathcal{S}} \{u_{i+1}(s) - u_i(s)\} - \min_{s \in \mathcal{S}} \{u_{i+1}(s) - u_i(s)\} < \epsilon$$

the greedy policy with respect to u_i is ϵ-optimal.

A.2   Concentration of reward.

Conditioned on the state-action counts $N(s, a)$, the rewards r_t are independent random variables. Hence, from concentration results for robust mean estimators, we have

$$P\left(\sum_{t=1}^{T} r_t \le \sum_{s,a} N(s, a)\bar{r}(s, a) - C_T \,\middle|\, (N(s, a))_{s,a}\right) \le \left(\frac{\delta}{8T}\right)^{5/4} < \frac{\delta}{12T^{5/4}} \tag{14}$$

where $C_T = (vT)^{\frac{1}{(1+\epsilon)}} \left(\frac{5c}{4} \log\left(\frac{8T}{\delta}\right)\right)^{\frac{\epsilon}{(1+\epsilon)}}$.

Letting $\Delta_k = \sum_{s,a} v_k(s, a)(\rho^* - \bar{r}(s, a))$ denote the regret for episode k, we can bound the total regret as:

$$\Delta(s_1, T) = T\rho^* - \sum_{t=1}^{T} r_t \le \sum_k \Delta_k + C_T \tag{15}$$

which holds with probability at least $1 - \frac{\delta}{12T^{5/4}}$.

A.3   Regret from failing confidence regions

We first bound the regret due to failing confidence regions. The idea is the same as that in Jaksch et al. [2010]: we show that since our confidence intervals are sufficiently large, the probability that they fail is very small per timestep.

The following lemma bounds the probability that $M \notin \mathcal{M}_k$:

Lemma 1. *For any $t \geq 1$,*

$$P\left(M \notin \mathcal{M}_t\right) < \frac{\delta}{15t^6} \tag{16}$$

Proof. This lemma and proof are essentially identical to Lemma 17 in [Jaksch et al., 2010]. Specifically, the probability that the confidence intervals in eqs. 7 and 2 are violated are bounded by $\frac{\delta}{60t^7 SA}$ and $\frac{\delta}{20t^7 SA}$ respectively. Then, a union bound over all timesteps and state-action pairs yields the desired result. □

Now, it follows identically from Jaksch et al. [2010] that $\sum_{k=1}^m \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} \leq \sqrt{T}$ with probability at least $1 - \frac{\delta}{12T^{5/4}}$. In our setting, since the rewards have bounded means, the regret in this scenario can be bounded by $R_\Delta \sqrt{T}$.

A.4   Regret for true MDP $M \in \mathcal{M}_k$.

We now bound the regret in each episode k assuming that our confidence set contains the true MDP.

In both UCRL2 and Heavy-UCRL2, the condition for stopping extended value iteration is identical. Hence, we similarly have by Theorem 7:

$$\Delta_k \leq \sum_{s,a} v_k(s,a)\left(\rho^* - \bar{r}(s,a)\right) \leq \sum_{s,a} v_k(s,a)\left(\widetilde{\rho}_k - \bar{r}(s,a)\right) + \sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}} \tag{17}$$

Following the notation of Jaksch et al. [2010], let $\widetilde{\mathbf{P}}_k$ denote the transition matrix of the optimistic MDP $\tilde{M}$, $\mathbf{P}_k$ the transition matrix of the true MDP, $\mathbf{v}_k$ the row vector corresponding to the visit counts for each state (and corresponding action chosen by $\widetilde{\pi}_k$), and $\mathbf{w}_k$ the vector defined as $w_k(s) := u_i(s) - \frac{\min_s u_i(s) + \max_s u_i(s)}{2}$. Then, the regret can be decomposed as follows:

$$\Delta_k \leq \mathbf{v}_k(\widetilde{\mathbf{P}}_k - \mathbf{P}_k)\mathbf{w}_k + \mathbf{v}_k(\mathbf{P}_k - \mathbf{I})\mathbf{w}_k + 2\sum_{s,a} v_k(s,a)v^{\frac{1}{1+\epsilon}}\left(\frac{7c\log\left(\frac{2SAt_k}{\delta}\right)}{\max\{1, N_k(s,a)\}}\right)^{\frac{\epsilon}{1+\epsilon}} + 2\sum_{s,a}\frac{v_k(s,a)}{\sqrt{t_k}} \tag{18}$$

Only the third term is dependent on the distribution of r. The rest of the terms hold identically as in Jaksch et al. [2010] with one small modification: for any iteration i in value iteration, the range of state values is bounded by $R_\Delta D$. To restate, the sum of the first and fourth terms is bounded by

$$R_\Delta D\left(\sqrt{14S\log\left(\frac{2AT}{\delta}\right)} + 2\right)\left(\sqrt{2} + 1\right)\sqrt{SAT} \tag{19}$$

and the second term is bounded by

$$R_\Delta\left(D\sqrt{\frac{5}{2}T\log\frac{8T}{\delta}} + DSA\log_2\left(\frac{8T}{SA}\right)\right) \tag{20}$$

To bound the third term, we require the following lemma, which generalizes Lemma 19 in Jaksch et al. [2010]:

Lemma 2. *For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$, there exists some constant C_ϵ such that*

$$\sum_{k=1}^n \frac{z_k}{Z_{k-1}^{\frac{\epsilon}{1+\epsilon}}} \leq C_\epsilon Z_n^{\frac{1}{1+\epsilon}}$$

Proof of Lemma 2. Let $C_\epsilon = \frac{2}{M}$, where $M = \sum_{k=1}^{\infty}\left(\frac{\frac{1}{1+\epsilon}}{k}\right) = 2^{\frac{1}{1+\epsilon}} - 1$. Note that $2 < C_\epsilon \leq 2(\sqrt{2} + 1)$.

We prove by induction on n. For the base case, we consider all n such that $\sum_{k=1}^{n-1} z_k \leq 1$. Then $Z_k = 1$ for all $k < n$ and $\sum_{k=1}^n \frac{z_k}{Z_{k-1}^{\frac{\epsilon}{1+\epsilon}}} = \sum_{k=1}^n z_k = \sum_{k=1}^{n-1} z_k + z_n \leq 1 + 1 < C_\epsilon Z_n^{\frac{1}{1+\epsilon}}$.

For the inductive step, assume the claim holds for some n. Then, we have

$$
\sum_{k=1}^{n+1} \frac{z_k}{Z_{k-1}^{\frac{\epsilon}{1+\epsilon}}} \leq C_\epsilon Z_n^{\frac{1}{1+\epsilon}} + \frac{z_{n+1}}{Z_n^{\frac{\epsilon}{1+\epsilon}}}
$$

$$
= \frac{1}{2} C_\epsilon \left(Z_n^{\frac{1}{1+\epsilon}} + \frac{z_{n+1}}{Z_n^{\frac{\epsilon}{1+\epsilon}}} \sum_{k=1}^{\infty} \binom{\frac{1}{1+\epsilon}}{k} \right) + \frac{1}{2} C_\epsilon Z_n^{\frac{1}{1+\epsilon}}
$$

$$
= \frac{1}{2} C_\epsilon \left(Z_n^{\frac{1}{1+\epsilon}} + \frac{1}{Z_n^{\frac{\epsilon}{1+\epsilon}}} \sum_{k=1}^{\infty} \binom{\frac{1}{1+\epsilon}}{k} z_{n+1}^{1-k} z_{n+1}^{k} \right) + \frac{1}{2} C_\epsilon Z_n^{\frac{1}{1+\epsilon}}
$$

$$
\leq \frac{1}{2} C_\epsilon \left(Z_n^{\frac{1}{1+\epsilon}} + \frac{1}{Z_n^{\frac{\epsilon}{1+\epsilon}}} \sum_{k=1}^{\infty} \binom{\frac{1}{1+\epsilon}}{k} Z_n^{1-k} z_{n+1}^{k} \right) + \frac{1}{2} C_\epsilon Z_n^{\frac{1}{1+\epsilon}}
$$

$$
\leq \frac{1}{2} C_\epsilon \left(Z_n^{\frac{1}{1+\epsilon}} + \sum_{k=1}^{\infty} \binom{\frac{1}{1+\epsilon}}{k} Z_n^{\frac{1}{1+\epsilon}-k} z_{n+1}^{k} \right) + \frac{1}{2} C_\epsilon Z_n^{\frac{1}{1+\epsilon}}
$$

$$
= \frac{1}{2} C_\epsilon \sum_{k=0}^{\infty} \left(\binom{\frac{1}{1+\epsilon}}{k} Z_n^{\frac{1}{1+\epsilon}-k} z_{n+1}^{k} \right) + \frac{1}{2} C_\epsilon Z_n^{\frac{1}{1+\epsilon}}
$$

$$
= \frac{1}{2} C_\epsilon (Z_n + z_{n+1})^{\frac{1}{1+\epsilon}} + \frac{1}{2} C_\epsilon Z_n^{\frac{1}{1+\epsilon}}
$$

$$
\leq C_\epsilon Z_{n+1}^{\frac{1}{1+\epsilon}}
$$

which completes the induction. $\qquad\square$

Via this lemma and Jensen's inequality, we have

$$
\sum_{s,a} \sum +k \frac{v_k(s,a)}{(\max\{1, N_k(s,a)\})^{\frac{\epsilon}{1+\epsilon}}} \leq C_\epsilon \sum_{s,a} N_{s,a}^{\frac{1}{1+\epsilon}} \leq C_\epsilon (SAT)^{\frac{1}{1+\epsilon}} \tag{21}
$$

which directly yields the following bound on the third term:

$$
2 C_\epsilon v^{\frac{1}{1+\epsilon}} \left(7c \log \left(\frac{2SAT}{\delta} \right) \right)^{\frac{\epsilon}{1+\epsilon}} (SAT)^{\frac{\epsilon}{1+\epsilon}} \tag{22}
$$

Combining all of the above yields the following bound on the regret:

$$
\Delta(T) \leq C_T + \sqrt{T} + R_\Delta \left(D \sqrt{\frac{5}{2} T \log \frac{8T}{\delta}} + DSA \log_2 \left(\frac{8T}{SA} \right) \right)
$$

$$
+ R_\Delta D \left(\sqrt{14S \log \left(\frac{2AT}{\delta} \right)} + 2 \right) \left(\sqrt{2} + 1 \right) \sqrt{SAT} \tag{23}
$$

$$
+ 2 C_\epsilon v^{\frac{1}{1+\epsilon}} \left(7c \log \left(\frac{2SAT}{\delta} \right) \right)^{\frac{\epsilon}{1+\epsilon}} (SAT)^{\frac{1}{1+\epsilon}}
$$

We can simplify the bound using similar manipulations as in Appendix C.4 of Jaksch et al. [2010]. Noting that we can bound the heavy tailed terms by

$$
(2 C_\epsilon + 1) v^{\frac{1}{1+\epsilon}} \left(7c \log \left(\frac{2SAT}{\delta} \right) \right)^{\frac{\epsilon}{1+\epsilon}} (SAT)^{\frac{1}{1+\epsilon}}
$$

Using $B_\epsilon := (2 C_\epsilon + 1) v^{\frac{1}{1+\epsilon}} (7c)^{\frac{\epsilon}{1+\epsilon}}$, we have the following bound on the regret:

$$
\Delta(s_0, T) \leq 20 R_\Delta DS \sqrt{AT \log \left(\frac{T}{\delta} \right)} + B_\epsilon \left(\log \left(\frac{2SAT}{\delta} \right) \right)^{\frac{\epsilon}{1+\epsilon}} (SAT)^{\frac{1}{1+\epsilon}} \tag{24}
$$

B Proof of Corollary 1

Proof. We simply have to compute T_0 such that

$$\frac{20R_\Delta DS\sqrt{AT\log\left(\frac{T}{\delta}\right)}}{T} + \frac{(2C_\epsilon+1)v^{\frac{1}{1+\epsilon}}(SAT)^{\frac{1}{1+\epsilon}}\left(7c\log\left(\frac{2SAT}{\delta}\right)\right)^{\frac{\epsilon}{1+\epsilon}}}{T} < \lambda$$

for all $T \geq T_0$.

Any T_0 satisfies the condition when both the first term and the second term are smaller than $\lambda/2$. The first term comes directly from Jaksch et al. [2010], replacing λ with $\lambda/2$. For the second term, the condition is equivalent to

$$T > \alpha\log\left(\frac{2SAT}{\delta}\right) = \alpha\log\left(\frac{2SA}{\delta}\right) + \alpha\log\left(\frac{T}{\delta}\right)$$

Using the fact that $x > 2\log x$ for $x > 0$, we see that $2\alpha\log\left(\frac{\alpha}{\delta}\right) > \alpha\log\left(\frac{T}{\delta}\right)$, from which the result follows immediately. $\qquad\square$

C Proof of Theorem 2

The proof of this theorem follows as a consequence of this lemma:

Lemma 3. *Call an episode λ-bad if its average regret is larger than λ. Let L_λ be the number of steps taken by Heavy-UCRL2 in λ-bad episodes up to step T. Then for any initial state $s \in \mathcal{S}$, any T satisfying*

$$T \geq \frac{20R_\Delta D^{\frac{2+2\epsilon}{1-\epsilon}}S^{\frac{2\epsilon}{1-\epsilon}}}{B_\epsilon A}, \tag{25}$$

and any $\lambda > 0$, with probability of at least $1 - 3\delta$,

$$L_\lambda(T) \leq 7c\log\left(\frac{2SAT}{\delta}\right)\left(\frac{4C_\epsilon+2}{\lambda}\right)^{\frac{1+\epsilon}{\epsilon}}(SA)^{\frac{1}{\epsilon}}$$

Proof of Lemma 3. The proof, like that of Theorems 4 and 11 in Jaksch et al. [2010], draws heavily on that of Theorem 1. Let K_λ denote the random set corresponding to the episodes that are λ-bad. Then, the regret of these episodes is bounded with probability at least $1 - 2\delta$ by

$$\Delta'_\lambda(s,T) \leq (vL_\lambda)^{\frac{1}{1+\epsilon}}\left(c\log\left(\frac{T}{\delta}\right)\right)^{\frac{\epsilon}{1+\epsilon}} + \sum_{k\in K_\lambda}\Delta_k\mathbb{1}M \in \mathcal{M}_k \tag{26}$$

We can leverage our previous bound on the regret within each episode (eq.18) with some small modifications. In particular, in this setting we need to bound $\sum_{k\in K_\epsilon}\sum_{s,a}\frac{v_k(s,a)}{\max\{1,N_k(s,a)\}^{\frac{\epsilon}{1+\epsilon}}}$. This is proven in lemma 4. Leveraging this result, we have

$$\Delta'_\epsilon(s,T) \leq (vL_\lambda)^{\frac{1}{1+\epsilon}}\left(c\log\left(\frac{T}{\delta}\right)\right)^{\frac{\epsilon}{1+\epsilon}} + R_\Delta D\left(\sqrt{14S\log\left(\frac{2AT}{\delta}\right)}\right)\sqrt{SAT}$$
$$+ \sum_{k\in K_\epsilon}\mathbf{v}_k(\mathbf{P}_k - \mathbf{I})\mathbf{w}_k\mathbb{1}_{M\in\mathcal{M}_k} + C_\epsilon(L_\lambda SA)^{\frac{1}{1+\epsilon}} \tag{27}$$

As before, the third term can be bounded by

$$R_\Delta\left(2\sqrt{L_\lambda\log\left(\frac{T}{\delta}\right)} + DSA\log_2\left(\frac{8T}{SA}\right)\right)$$

So, we can simplify

$$\Delta'_\lambda(s,T) \le 20DS\sqrt{L_\lambda A \log\left(\frac{T}{\delta}\right)} + B_\epsilon \left(\log\left(\frac{2SAT}{\delta}\right)\right)^{\frac{\epsilon}{1+\epsilon}} (L_\lambda SA)^{\frac{1}{1+\epsilon}} \tag{28}$$

Under the regime given by eq. 25, this bound is equivalent to $2B_\epsilon \left(\log\left(\frac{2SAT}{\delta}\right)\right)^{\frac{\epsilon}{1+\epsilon}} (L_\lambda SA)^{\frac{1}{1+\epsilon}}$. Finally, using the fact that $\lambda L_\lambda \le \Delta'_\lambda(s,T)$ yields the desired bound.

$\square$

We now prove the following lemma:

Lemma 4.

$$\sum_{k \in K_\lambda} \sum_{s,a} \frac{v_k(s,a)}{\max\{1, N_k(s,a)\}^{\frac{\epsilon}{1+\epsilon}}} \le C_\epsilon (L_\lambda SA)^{\frac{1}{1+\epsilon}}$$

Proof. The proof is identical to that of eq. 27 in Jaksch et al. [2010]. The only modification is that we have $d_k := \max\{1, N_k(s,a)\}^{\frac{\epsilon}{1+\epsilon}}$. Since the proof only uses the property that $d_k \le d_l$ for $k \le l$, it follows that we have

$$\sum_{k=1}^m \frac{v_k}{d_k} \mathbb{1}_{k \in K_\lambda} \le \sum_{k=1}^{m_\lambda} \frac{v'_k}{d_k} \le C_\epsilon \ell_\lambda^{\frac{1}{1+\epsilon}}$$

Summing over all state-action pairs and applying Jensen's inequality yields the desired result. $\square$

Finally, we are ready to prove the main theorem.

Proof of Theorem 2. The first part of the theorem follows immediately from the preceding lemma. For the second part of the theorem, note that the expected regret in $\frac{g}{2}$-bad episodes is upper bounded by $7c\log\left(\frac{2SAT}{\delta}\right)(4C_\epsilon + 2)^{\frac{1+\epsilon}{\epsilon}} \left(\frac{2SA}{g}\right)^{1/\epsilon} + 1$. The theorem statement then follows by identical argument as in Jaksch et al. [2010]. $\square$

D   Proof of Theorem 3

Proof of Theorem 3. We first describe the intuition behind the proof. Essentially, if T is sufficiently large, the regret of the Heavy-UCRL2 is $\tilde{O}\left((SAT)^{\frac{1}{1+\epsilon}}\right)$. Hence, restarting Heavy-UCRL2 every $\left(\frac{T}{\ell}\right)^{\frac{1+\epsilon}{1+2\epsilon}}$ steps, the regret for the ℓ changing MDP periods is bounded by $R_\Delta \ell^{\frac{\epsilon}{1+2\epsilon}} T^{\frac{1+\epsilon}{1+2\epsilon}}$. Furthermore, since we restart Heavy-UCRL2 $T^{\frac{\epsilon}{1+2\epsilon}} \ell^{\frac{1+\epsilon}{1+2\epsilon}}$ times, the regret incurred in non-changing stages is also $\tilde{O}\left(R_\Delta \ell^{\frac{\epsilon}{1+2\epsilon}} T^{\frac{1+\epsilon}{1+2\epsilon}}\right)$. Thus, the total regret is bounded by $\tilde{O}\left(R_\Delta \ell^{\frac{\epsilon}{1+2\epsilon}} T^{\frac{1+\epsilon}{1+2\epsilon}}\right)$.

Since the time horizon T is not assumed to be known a priori, we use a variant of the "doubling trick." Specifically, we restart Heavy-UCRL2 with parameter $\frac{\delta}{\ell^2}$ at steps $\left\lceil \frac{i^{(1+2\epsilon)/\epsilon}}{\ell^{(1+\epsilon)/\epsilon}} \right\rceil$ for $i = 1, 2, 3, \ldots$, effectively dividing the algorithm into n stages. For some fixed time horizon T, the total number of restarts n is bounded by

$$\ell^{\frac{1+\epsilon}{1+2\epsilon}} T^{\frac{\epsilon}{1+2\epsilon}} - 1 \le n \le \ell^{\frac{1+\epsilon}{1+2\epsilon}} T^{\frac{\epsilon}{1+2\epsilon}} \tag{29}$$

The regret Δ_r incurred in the ℓ stages in which the MDP is restarted is bounded by R_Δ times the total number of steps in these stages. This latter value is maximized when they occur in the last ℓ stages, which occur T_ℓ

timesteps. Then we have

$$
\begin{aligned}
T_\ell &\leq \frac{1}{\ell^{(1+\epsilon)/\epsilon}} \left((n+1)^{(1+2\epsilon)/\epsilon} - (n-\ell+1)^{(1+2\epsilon)/\epsilon} \right) \\
&= \frac{1}{\ell^{(1+\epsilon)/\epsilon}} \sum_{k=0}^{\infty} \binom{\frac{1+2\epsilon}{\epsilon}}{k} n^{\frac{1+2\epsilon}{\epsilon}-k} \left(1 - (1-\ell)^k \right) \\
&\leq \frac{1+2\epsilon}{\epsilon} n^{\frac{1+\epsilon}{\epsilon}} \ell^{-\frac{1}{\epsilon}} \\
&\leq \frac{1+2\epsilon}{\epsilon} \ell^{\frac{\epsilon}{1+2\epsilon}} T^{\frac{1+\epsilon}{1+2\epsilon}}
\end{aligned}
\tag{30}
$$

where in the second, third, and fourth (in)equalities we used the generalized binomial theorem, that the sum is dominated by $k = 1$, and eq. 29 respectively. The total regret Δ_r is then bounded by $R_\Delta \ell T_\ell = R_\Delta \frac{1+2\epsilon}{\epsilon} \ell^{\frac{1+\epsilon}{1+2\epsilon}} T^{\frac{1+\epsilon}{1+2\epsilon}}$.

We now consider the regret for the stages in which the MDP does not change. Letting $T_i := \min(T, \tau_{i+1}) - \tau_i$, we have by Theorem 1 that

$$
\begin{aligned}
\Delta(S_{\tau_i, T_i}) &\leq 2B_\epsilon \left(\log \left(\frac{2\ell^2 SAT_i}{\delta} \right) \right)^{\frac{\epsilon}{1+\epsilon}} (SAT_i)^{\frac{1}{1+\epsilon}} \\
&\leq 2B_\epsilon \left(3 \log \left(\frac{2SAT}{\delta} \right) \right)^{\frac{\epsilon}{1+\epsilon}} (SAT_i)^{\frac{1}{1+\epsilon}}
\end{aligned}
\tag{31}
$$

with probability $\frac{\delta}{4\ell^2 T_i^{5/4}}$.

Summing over all stages $i = 1, \ldots, n$, the total regret Δ_f is bounded by

$$
\begin{aligned}
\Delta_f &= \sum_{i=1}^{n} \Delta(S_{\tau_i, T_i}) \\
&\leq 2B_\epsilon \left(3n \log \left(\frac{2SAT}{\delta} \right) \right)^{\frac{\epsilon}{1+\epsilon}} (SAT)^{\frac{1}{1+\epsilon}} \\
&\leq 2B_\epsilon \ell^{\frac{\epsilon}{1+2\epsilon}} T^{\frac{1+\epsilon}{1+2\epsilon}} \left(3 \log \left(\frac{2SAT}{\delta} \right) \right)^{\frac{\epsilon}{1+\epsilon}} (SA)^{\frac{1}{1+\epsilon}},
\end{aligned}
\tag{32}
$$

where the first inequality is due to Jensen's inequality. Finally, it follows from Jaksch et al. [2010] that the total probability is bounded by $1 - \delta$. $\qquad \square$

E   Proof of Theorem 4

Our proof follows that of Theorem 1 in Jin et al. [2018]. We first prove that $Q^k - Q^*$ is bounded for all s, a, h, k, then use that fact to recursively decompose the regret. We also use the similar notation: s_h^k and a_h^k denote the state and action taken at step h in episode k, and define

$$
\alpha_t^0 = \prod_{j=1}^{t} (1 - \alpha_j), \qquad \alpha_t^i = \alpha_i \prod_{j=i+1}^{t} (1 - \alpha_j)
\tag{33}
$$

We consider the expected regret, and for clarity we omit the expectation over the reward stochasticity.

We first prove a useful auxiliary lemma about the learning rate, which generalizes Lemma 4.1.a in Jin et al. [2018]:

Lemma 5. *For all α_t^i, we have the following:*

$$
t^{-\frac{\epsilon}{1+\epsilon}} \leq \sum_{i=1}^{t} \alpha_t^i i^{-\frac{\epsilon}{1+\epsilon}} \leq 2t^{-\frac{\epsilon}{1+\epsilon}}
\tag{34}
$$

Proof. We prove via induction on t. For the base case, $t = 1$, and $\alpha_1^1 = 1$. Now assume the hypothesis holds for

$t = 1, \ldots, k - 1$. For the lower bound, we have:

$$
\begin{aligned}
\sum_{i=1}^{k} \alpha_k^i i^{-\frac{\epsilon}{1+\epsilon}} &= \alpha_k k^{-\frac{\epsilon}{1+\epsilon}} + (1 - \alpha_k) \sum_{i=1}^{k-1} \alpha_{k-1}^i i^{-\frac{\epsilon}{1+\epsilon}} \\
&\geq \alpha_k k^{-\frac{\epsilon}{1+\epsilon}} + (1 - \alpha_k)(k-1)^{-\frac{\epsilon}{1+\epsilon}} \\
&\geq \alpha_k k^{-\frac{\epsilon}{1+\epsilon}} + (1 - \alpha_k) k^{-\frac{\epsilon}{1+\epsilon}} \\
&= k^{-\frac{\epsilon}{1+\epsilon}}
\end{aligned}
$$

and for the upper bound:

$$
\begin{aligned}
\sum_{i=1}^{k} \alpha_k^i i^{-\frac{\epsilon}{1+\epsilon}} &= \alpha_k k^{-\frac{\epsilon}{1+\epsilon}} + (1 - \alpha_k) \sum_{i=1}^{k-1} \alpha_{k-1}^i i^{-\frac{\epsilon}{1+\epsilon}} \\
&\leq \alpha_k k^{-\frac{\epsilon}{1+\epsilon}} + 2(1 - \alpha_k)(k-1)^{-\frac{\epsilon}{1+\epsilon}} \\
&= \frac{H+1}{H+k} k^{-\frac{\epsilon}{1+\epsilon}} + \frac{2k^{\frac{1}{1+\epsilon}}}{H+k} \\
&= 2k^{-\frac{\epsilon}{1+\epsilon}} - \frac{H-1}{H+k} k^{-\frac{\epsilon}{1+\epsilon}} \\
&\leq 2k^{-\frac{\epsilon}{1+\epsilon}}
\end{aligned}
$$

which completes the induction. $\square$

The following lemma, corresponding to lemmas 4.3 and C.4 in Jin et al. [2018], upper and lower bounds the gap between the estimated Q-value function at each episode and and Q^*.

Lemma 6. *Let* $\beta_t' = 2 \sum_{i=1}^{t} \alpha_t^i b_i' \leq 16 c r_{max} \sqrt{H^3 \iota / t} + 8 H u^{\frac{1}{1+\epsilon}} (\iota/t)^{\frac{\epsilon}{1+\epsilon}}$ *(by lemma 5). Then for all* (s, a, h, k), $(Q_h^k - Q_h^*)(s, a)$ *satisfies the following bound with probability at least* $1 - \delta$:

$$
0 \leq (Q_h^k - Q_h^*)(s, a) \leq \alpha_t^0 H r_{max} + \sum_{i=1}^{T} \alpha_t^i (V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i}) + \beta_t' \tag{35}
$$

Proof. Since the reward is stochastic, the following identity holds for Q^*:

$$
Q_h^*(s, a) = \alpha_t^0 Q_h^*(s, a) + \sum_{i=1}^{t} \alpha_t^i \left[\bar{r}_h(s, a) + (\mathbb{P}_h - \hat{\mathbb{P}}_h^{k_i}) V_{h+1}^*(s, a) + V_{h+1}^*(s_{h+1}^{k_i}) \right] \tag{36}
$$

Hence, we have

$$
\begin{aligned}
(Q_h^k - Q_h^*)(x, a) &\leq \alpha_t^0 (H r_{max} - Q_h^*(s, a)) + \sum_{i=1}^{t} \alpha_t^i \left[(V_{h+1}^{k_i} - V_{h+1}^*)(s_{h+1}^{k_i}) + [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) V_{h+1}^*](s, a) + b_i \right] \\
&\quad + \sum_{i=1}^{t} \alpha_t^i \left(r_h(s, a) \mathbb{1}_{|r_h(s,a)| \leq B_t} - \bar{r}_h(s, a) \right)
\end{aligned} \tag{37}
$$

We first consider the last term. Since $\alpha_t^i \leq \frac{2H}{t}$, this quantity is bounded by $\frac{2H}{t} \sum_{i=1}^{t} (r_h \mathbb{1}_{|r_h(s,a)| \leq B_t} - \bar{r}_h)(x, a)$. By Lemma 1 from Bubeck et al. [2013], this is upper (and lower) bounded by $\pm 8 H u^{\frac{1}{1+\epsilon}} \left(\frac{\iota}{t} \right)^{\frac{\epsilon}{1+\epsilon}}$ with probability at least $1 - \frac{\delta}{2SAT}$. A union bound over all s, a, t implies that this uniformly holds with probability at least $1 - \frac{\delta}{2}$.

It remains to modify the computation of the second term. Here, the setup is almost identical: we let $k_i := \min \left(\{ k \in [K] \mid k > k_{i-1} \wedge (s_h^k, a_h^k) = (s, a) \} \bigcup \{ K + 1 \} \right)$. Then $(\mathbb{1}[k_i \leq K] \cdot [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) V_{h+1}^*](s, a))_{i=1}^{\tau}$ is a martingale difference sequence with respect to the filtration consisting of the σ-field generated by all random variables up to step h in episode k_i. The main difference in this setting is that V^* is now uniformly bounded by

Hr_{max}. Hence, a straightforward application of Azuma-Hoeffding followed by a union bound over $k_i \leq K$ results in the following bound, which holds with probability $1 - \delta/SAH$:

$$\sum_{i=1}^{\tau} \alpha_{\tau}^i \cdot \mathbb{1}[k_i \leq K] \cdot [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)V_{h+1}^*](s,a) \leq cHr_{\max}\sqrt{\sum_{i=1}^{\tau}(2\alpha_{\tau}^i)^2 \cdot \iota} \leq cr_{\max}\sqrt{\frac{H^3\iota}{\tau}} \tag{38}$$

$\square$

We are now ready to prove theorem 4.

Proof. The proof follows identically as that of Theorem 2 in Jin et al. [2018] except for the following two differences: 1. β_t' has an additional additive term, and 2. the martingale difference sequence involving $V^* - V^k$ is now bounded by Hr_{max} instead of merely H. We compute the additional regret incurred for each term:

- **Increase in** $\sum_{h=1}^{H}\sum_{k=1}^{K}\beta_{n_h^k}$. The additional regret incurred by β' relative to β is given by

$$H^2 u^{\frac{1}{1+\epsilon}}\iota^{\frac{\epsilon}{1+\epsilon}}\sum_{s,a}\sum_{n=1}^{N_h^K(s,a)} n^{-\frac{\epsilon}{1+\epsilon}} \leq \Theta\left(H^2(SA\iota)^{\frac{\epsilon}{1+\epsilon}}K^{\frac{1}{1+\epsilon}}\right) \leq \Theta\left(H(SA\iota)^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}\right)$$

  because the sum is maximized when $N_h^K(s,a) = \frac{K}{SA}$ for all s,a.

- **Regret scaling due to** $V^* - V^k$. By Azuma-Hoeffding, the regret is scaled by a r_{max} factor, i.e. $cHr_{max}\sqrt{T\iota}$. This term is strictly dominated by the previous one.

Hence, the total regret is the same as that in Jin et al. [2018] scaled by r_{max} with an additional $\Theta\left(H^2(SA\iota)^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}\right)$ factor. $\square$

E.1 Proof of Theorem 5

Theorem 5. *In the finite-horizon episodic MDP setting, the regret of Heavy-Q-Learning with UCB-Bernstein is*

$$\Theta\left(\sqrt{H^3 r_{max}^3 SAT\iota} + \sqrt{H^{\frac{1+4\epsilon}{\epsilon}}r_{max}S^2 A^2\iota^2} + \sqrt{H^9 r_{max}^2 u^{\frac{1}{1+\epsilon}}S^3 A^3\iota^3} + H^2(SA\iota)^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}} + H^{\frac{1+3\epsilon}{\epsilon}}\sqrt{S^3 A^3\iota^4\epsilon}\right)$$

.

Proof of Theorem 5. We first re-define notation from Jin et al. [2018]. The variance operator for the next-state value is given by

$$[\mathbb{V}_h V_{h+1}](s,a) := \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}\left[V_{h+1}(s') - [\mathbb{P}_h V_{h+1}](s,a)\right]^2 \tag{39}$$

and the empirical variance estimate

$$W_t(s,a,h) := \frac{1}{t}\sum_{i=1}^{t}\left[V_{h+1}^{k_i}(s_{h+1}^{k_i}) - \frac{1}{t}\sum_{j=1}^{t}V_{h+1}^{k_j}(s_{h+1}^{k_j})\right]^2 \tag{40}$$

To define the bonus, Jin et al. [2018] define

$$\beta_t := \min\left\{c_1\left(\sqrt{\frac{H}{t}\cdot(W_t(s,a,h)+H)\iota} + \frac{\sqrt{H^7 SA \cdot \iota}}{t}\right), c_2\sqrt{\frac{H^3\iota}{t}}\right\} \tag{41}$$

and let

$$b_1(s,a,h) := \frac{\beta_1(s,a,h)}{2}, \qquad b_t(s,a,h) := \frac{\beta_t(s,a,h) - (1-\alpha_t)\beta_{t-1}(s,a,h)}{2\alpha_t} \tag{42}$$

We claim that the proof of Theorem 2 in Jin et al. [2018] holds in the stochastic heavy-tailed reward setting when using the following bonus:

$$\beta'_t := \min\left\{ c_1 \left(\sqrt{\frac{Hr_{max}}{t} \cdot (W_t(s,a,h) + H)\iota} + \frac{\sqrt{H^7 r_{max} SA} \cdot \iota}{t} + \frac{H^2 \iota \sqrt{r_{max}^3}}{t} + \frac{H^{\frac{1+2\epsilon}{\epsilon}} \iota \sqrt{SA\epsilon}}{t} \right), c_2 r_{max} \sqrt{\frac{H^3 \iota}{t}} \right\}$$
$$+ 8Hu^{\frac{1}{1+\epsilon}}(\iota/t)^{\frac{\epsilon}{1+\epsilon}}$$

$$(43)$$

We proceed to calculate the regret following the steps of Jin et al. [2018]. Again, we use the same notation $\phi_h^k = (V_h^k - V_h^*)(s_h^k)$ and $\delta_h^k = (V_h^k - V_h^{\pi_k})(s_h^k)$.

Coarse upper bound on $Q^k - Q^*$. Note that that the corresponding bound for this term under UCB-Hoeffding (eq. 35) holds here identically. This fact will be used in the following to bound the gap between W_t and $\mathbb{V}_h V_{h+1}$.

Bounding the gap between W_t and $\mathbb{V}_h V_{h+1}$ We first translate lemma C.7 Jin et al. [2018] to our setting, showing that for some non-negative weight vector $w = (w_1, \ldots, w_k)$, the sum $\sum_{k=1}^K \phi_h^k$ is bounded by $\sum_{k=1}^K (Q_h^k - Q_h^*)(s_h^k, a_h^k)$. In the bounded $[0,1]$ reward setting, this bound can be computed as $O(SA\|w\|_\infty \sqrt{H^5 \iota} + \sqrt{SA\|w\|_1 \|w\|_\infty H^5 \iota})$. In our setting, these terms are scaled by r_{max}, plus an additional term due to our modified β'_t. The increase due to this term is given by

$$O(H) \cdot \sum_{k=1}^K w_k H u^{\frac{1}{1+\epsilon}} \left(\frac{\iota}{n_h^k}\right)^{\frac{\epsilon}{1+\epsilon}} = O(H) \cdot \sum_{s,a} \sum_{i=1}^{N_h^K(s,a)} w_{k_i(s,a)} H u^{\frac{1}{1+\epsilon}} \left(\frac{\iota}{n_h^k}\right)^{\frac{\epsilon}{1+\epsilon}}$$

$$\leq \Theta\left(H^2 u^{\frac{1}{1+\epsilon}} \iota^{\frac{\epsilon}{1+\epsilon}}\right) \cdot \sum_{s,a} \|w\|_\infty \left(1 + \sum_{i=1}^{\left\lfloor \frac{\|w\|_1}{SA\|w\|_\infty} \right\rfloor} \left(\frac{1}{i}\right)^{\frac{\epsilon}{1+\epsilon}}\right)$$

$$(44)$$

$$\leq \Theta\left(H^2 u^{\frac{1}{1+\epsilon}} \iota^{\frac{\epsilon}{1+\epsilon}} \left(SA\|w\|_\infty + (SA\|w\|_\infty)^{\frac{\epsilon}{1+\epsilon}} \|w\|_1^{\frac{1}{1+\epsilon}}\right)\right)$$

We can now apply this lemma to compute the increase in the bound between the empirical and actual variance. Recall that Jin et al. [2018] bound the gap between $P_1 := [\mathbb{V}_h V_{h+1}](s,a)$ and $P_4 := W_t(s,a,h)$ via the triangle inequality on intermediate terms

$$P_2 := \frac{1}{t} \sum_{i=1}^t \left[V_{h+1}^*(s_{h+1}^{k_i}) - [\mathbb{P}V_{h+1}^*](s,a) \right]^2$$

$$P_3 := \frac{1}{t} \sum_{i=1}^t \left[V_{h+1}^*(s_{h+1}^{k_i}) - \frac{1}{t} \sum_{j=1}^t V_{h+1}^*(s_{h+1}^{k_j}) \right]^2$$

$$(45)$$

We proceed to analyze the gaps $|P_1 - P_2|$, $|P_2 - P_3|$, $|P_3 - P_4|$ in this setting. For $|P_1 - P_2|$ and $|P_2 - P_3|$, straightforward applications of the Azuma-Hoeffding inequality show that they are bounded by $cH^2 r_{max}^2 \sqrt{\iota/t}$. For $|P_3 - P_4|$, we apply our preceding lemma: taking w so that $w_{k_i} = \frac{1}{t}$ for $i = 1, \ldots, t$ and 0 otherwise, so that $\|w\|_1 = 1$ and $\|w\|_\infty = \frac{1}{t}$, yields

$$|P_3 - P_4| \leq O\left(r_{max} \sqrt{H^7 \iota} \left(\frac{SA}{t} + \sqrt{\frac{SA}{t}} \right) + \frac{H^3 u^{\frac{1}{1+\epsilon}} \iota^{\frac{\epsilon}{1+\epsilon}} SA}{t} + \frac{H^3 u^{\frac{1}{1+\epsilon}} (SA\iota)^{\frac{\epsilon}{1+\epsilon}}}{t^{\frac{\epsilon}{1+\epsilon}}} \right)$$

$$(46)$$

Hence, the gap between the empirical variance and the actual variance for Heavy-Q-Learning with UCB-Bernstein is upper bounded by

$$\Theta\left(H^2 r_{max}^2 \sqrt{\frac{\iota}{t}} + r_{max} \sqrt{H^7 \iota} \left(\frac{SA}{t} + \sqrt{\frac{SA}{t}} \right) + \frac{H^3 u^{\frac{1}{1+\epsilon}} \iota^{\frac{\epsilon}{1+\epsilon}} SA}{t} + \frac{H^3 u^{\frac{1}{1+\epsilon}} (SA\iota)^{\frac{\epsilon}{1+\epsilon}}}{t^{\frac{\epsilon}{1+\epsilon}}} \right)$$

$$(47)$$

Applying Freedman's inequality. Recall that for UCB-Bernstein, an application of Freedman's inequality allows us to bound the martingale difference sequence in eq. 38 by $O\left(\sqrt{\frac{H}{t}[\mathbb{V}_h V_{h+1}^*]\iota + \frac{H^2}{t}r_{max}\iota}\right)$ [Freedman, 1975]. Jin et al. [2018] show that in the bounded $[0, 1]$ reward setting, this term is bounded by

$$\Theta\left(\sqrt{\frac{H}{t}(W_t(s, a, h) + H)\iota} + \frac{\iota\sqrt{H^7 SA}}{t}\right) \tag{48}$$

In the following, we modify the computation of this bound using eq. 47:

$$\Theta\left(\sqrt{\frac{H\iota}{t}\left(W_t(s, a, h)r_{max} + Hr_{max} + H^2 r_{max}^2\sqrt{\frac{\iota}{t}} + H^3 u^{\frac{1}{1+\epsilon}}\iota^{\frac{\epsilon}{1+\epsilon}}\left(\frac{SA}{t} + \left(\frac{SA\iota}{t}\right)^{\frac{\epsilon}{1+\epsilon}}\right)\right)} + \frac{\iota\sqrt{H^7 SA r_{max}}}{t}\right)$$

$$\leq \Theta\left(\sqrt{\frac{H\iota}{t}\left(W_t(s, a, h)r_{max} + Hr_{max} + \frac{1}{1+\epsilon}\left(H + \epsilon\frac{H^{\frac{2+3\epsilon}{\epsilon}}SA\iota}{t}\right)\right)} + \frac{H^2\iota\sqrt{r_{max}^3}}{t} + \frac{\iota\sqrt{H^7 SA r_{max}}}{t}\right)$$

$$\leq \Theta\left(\sqrt{\frac{Hr_{max}\iota}{t}(W_t(s, a, h) + H)} + \frac{H^{\frac{1+2\epsilon}{\epsilon}}\iota\sqrt{SA\epsilon}}{t} + \frac{H^2\iota\sqrt{r_{max}^3}}{t} + \frac{\iota\sqrt{H^7 SA r_{max}}}{t}\right) \leq \beta_t \tag{49}$$

where we repeatedly leverage the (weighted) AM-GM inequality.

Analogue of Lemma C.5. We now show that the analysis of the total variance introduces an unavoidable r_{max}^2 factor. The analysis follows identically as that of Lemma C.5 in Jin et al. [2018], except noting that V is now bounded by Hr_{max}, which implies that the total variance is bounded by $H^2 r_{max}^2$. Namely, we have the following:

$$\sum_{k=1}^K \sum_{h=1}^H \mathbb{V}_h V_{h+1}^{\pi_k} \leq O\left(r_{max}^2\left(HT + H^3\iota\right)\right) \tag{50}$$

Analogue of Lemma C.6. We also state the following analogue of Lemma C.6 in Jin et al. [2018], upper bounding W_t via eq. 50. The proof of this lemma is a straightforward modification of that for the bounded reward setting.

Lemma 7. *There exists constant c such that the following holds with probability $1 - 4\delta$*

$$W_t(s, a, h) \leq \mathbb{V}_h V_{h+1}^{\pi_h^k}(s, a) + 2Hr_{max}(\delta_{h+1}^k + \xi_{h+1}^k)$$
$$+ c\left(H^2 r_{max}^2\sqrt{\frac{\iota}{t}} + r_{max}\sqrt{H^7\iota}\left(\frac{SA}{t} + \sqrt{\frac{SA}{t}}\right) + H^3 u^{\frac{1}{1+\epsilon}}\left(\frac{\iota^{\frac{\epsilon}{1+\epsilon}}SA}{t} + \frac{(SA\iota)^{\frac{\epsilon}{1+\epsilon}}}{t^{\frac{\epsilon}{1+\epsilon}}}\right)\right)$$

Computing total regret. Here, we show the necessary modifications to the computation of regret akin to those in Theorem 2 Jin et al. [2018]. We first re-state the recursive expansion of the regret:

$$\sum_{k=1}^K \delta_h^k \leq SAH^2 + \sum_{h'=h}^H \sum_{k=1}^K (\beta_{n_{h'}^k}(s_{h'}^k, a_{h'}^k, h') + \xi_{h+1}^k) \tag{51}$$

where $\xi_{h+1}^k := [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h)(V_{h+1}^* - V_{h+1}^k)](s, a)$. By Azuma-Hoeffding, we have with probability $1 - \delta$

$$\sum_{h'=h}^H \sum_{k=1}^K \xi_{h'}^k \leq \Theta(Hr_{max}\sqrt{T\iota}) \tag{52}$$

We also recall eq. C.13 from Jin et al. [2018], scaled by a r_{max} factor in our setting:

$$\sum_{k=1}^K \delta_h^k \leq \Theta(r_{max}\sqrt{H^4 SAT\iota}) \tag{53}$$

We now proceed to bound $\sum_{k=1}^{K} \sum_{h=1}^{H} \beta_{n_h^k}$:

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \beta_{n_h^k} \leq \sum_{k=1}^{K} \sum_{h=1}^{H} \Theta\left(\sqrt{\frac{Hr_{max}}{t}(W_t(s,a,h)+H)\iota} + \frac{\iota\sqrt{H^7 r_{max} SA}}{t} + \frac{H^2\iota\sqrt{r_{max}^3}}{t} + \frac{H^{\frac{1+2\epsilon}{\epsilon}}\iota\sqrt{SA\epsilon}}{t}\right) +$$
$$\sum_{k=1}^{K} \sum_{h=1}^{H} Hu^{\frac{1}{1+\epsilon}}(\iota/t)^{\frac{\epsilon}{1+\epsilon}} \tag{54}$$

The second sum is bounded similarly as in the analysis of UCB-Hoeffding:

$$\sum_{k=1}^{K} \sum_{h=1}^{H} Hu^{\frac{1}{1+\epsilon}}(\iota/t)^{\frac{\epsilon}{1+\epsilon}} \leq \Theta\left(H^2(SA\iota)^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}\right) \tag{55}$$

We now bound the second, third, and fourth terms in the first sum using the fact that $\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{n_h^k} \leq HSA\iota$:

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{\iota\sqrt{H^7 r_{max} SA}}{t} \leq \sqrt{H^9 r_{max} S^3 A^3 \iota^4}$$
$$\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{H^2\iota\sqrt{r_{max}^3}}{t} \leq H^3 SA\iota^2 \sqrt{r_{max}^3} \tag{56}$$
$$\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{H^{\frac{1+2\epsilon}{\epsilon}}\iota\sqrt{SA\epsilon}}{t} \leq H^{\frac{1+3\epsilon}{\epsilon}}\sqrt{S^3 A^3 \iota^4 \epsilon}$$

We now bound $\sum_{k=1}^{K} \sum_{h=1}^{H} \sqrt{\frac{Hr_{max}}{t}(W_t(s,a,h)+H)\iota}$. From eq. C.15 in Jin et al. [2018], this is upper bounded by $\sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} W_t(s,a,h)} \cdot \sqrt{H^2 r_{max} SA\iota} + \sqrt{H^3 r_{max} SAT\iota}$.

We now have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} W_{n_h^k}(s,a,h) \leq \sum_{k=1}^{K} \sum_{h=1}^{H}\left(\mathbb{V}_h V_{h+1}^{\pi_h^k}(s_h^k, a_h^k) + 2Hr_{max}(\delta_{h+1}^k + \xi_{h+1}^k) + \right.$$
$$\left. \sum_{k=1}^{K} \sum_{h=1}^{H} \Theta\left(H^2 r_{max}^2 \sqrt{\frac{\iota}{t}} + r_{max}\sqrt{H^7\iota}\left(\frac{SA}{t} + \sqrt{\frac{SA}{t}}\right) + H^3 u^{\frac{1}{1+\epsilon}}\left(\frac{\iota^{\frac{\epsilon}{1+\epsilon}}SA}{t} + \frac{(SA\iota)^{\frac{\epsilon}{1+\epsilon}}}{t^{\frac{\epsilon}{1+\epsilon}}}\right)\right)\right) \tag{57}$$

Substituting equations 50, 52, and 53, this is bounded by

$$\Theta\left(r_{max}^2 \sqrt{H^8 SAT\iota} + Hr_{max}^2 T + H^3 r_{max}^2 \iota + r_{max} S^2 A^2 \sqrt{H^9 \iota^3} + \right.$$
$$\left. r_{max} SA\sqrt{H^8 T\iota} + H^3 u^{\frac{1}{1+\epsilon}}\iota^{\frac{1+2\epsilon}{1+\epsilon}}S^2 A^2 + H^{\frac{2+3\epsilon}{1+\epsilon}}u^{\frac{1}{1+\epsilon}}(S^2 A^2 \iota)^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}\right) \tag{58}$$

Noting that $\epsilon^2 \leq 1 + \epsilon$ and $r_{max} \leq u^{\frac{1}{1+\epsilon}}$ by Cauchy-Schwarz, we can simplify this bound into the following:

$$\sum_{k=1}^{K} \sum_{h=1}^{H} W_{n_h^k}(s,a,h) \leq \Theta\left(Hr_{max}^2 T + r_{max}^2 S^2 A^2 H^7 \iota + H^{\frac{1+2\epsilon}{1+\epsilon}}SA\iota + u^{\frac{1}{1+\epsilon}}S^2 A^2 \sqrt{H^9\iota^3}\right) \tag{59}$$

Hence, we have the following:

$$\sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} W_{n_h^k}(s,a,h)} \cdot \sqrt{H^2 r_{max} SA\iota} \leq \Theta\left(\sqrt{H^3 r_{max}^3 SAT\iota} + \sqrt{H^{\frac{1+4\epsilon}{\epsilon}}r_{max}S^2 A^2 \iota^2} + \sqrt{H^9 r_{max}^2 u^{\frac{1}{1+\epsilon}}S^3 A^3 \iota^3}\right) \tag{60}$$

Substituting everything back into eqs. 51 and 54 yields the desired result.

$\square$

F   Experiment Details

F.1   Synthetic MDPs

Here, we describe the SixArms and DoubleChain MDPs in further detail. SixArms consists of six states $s_1, \ldots, s_6$ rearranged around a central initial state s_0 [Strehl and Littman, 2008]. There are six actions, and only taking action $i - 1$ in state i for $i = 1, \ldots, 6$ yields positive reward. Taking action i in state s_0 transitions to state i with probability p_i, and taking action $i - 1$ state s_i, $i > 0$ transitions to itself with probability 1, and taking any action transitions back to s_0. Letting r_i denote $r(i, i - 1)$, we set r_1 to $\mathcal{N}(1.20, 0.1^2)$, and r_i to $\mathcal{L}(\mu_i, 1.1, 0, 1)$ where $\mu_i = 1 + 0.2(i - 1)$ for $i = 1, \ldots, 6$. On SixArms, we run all algorithms for $5 \cdot 10^5$ episodes of length 25 for 30 random seeds.

DoubleChain, pictured in figure 2, begins at an initial state s_0, and can transition to either of two RiverSwim-style MDPs of length l [Dimakopoulou and Van Roy, 2018]. As in RiverSwim, only the rightmost states have significant positive reward; all other states have either zero or slightly negative mean reward. In our DoubleChain experiments, the reward distributions for the two rightmost states s_l and s_{2l+1} are set to $\mathcal{N}(0.5, 0.1^2)$ and symmetric ($\beta = 0$) Levy α-stable with $\alpha = 1.1$, $\mu = 1$ respectively. The rewards for states $1, \ldots, l - 1$ are set to $\mathcal{N}(0, 0.1^2)$, and states $l + 1, \ldots, 2l$ to $\mathcal{N}(-0.1, 0.01^2)$. We set $p = 0.8$, $l = 3$, and run all algorithms for $5 \cdot 10^5$ episodes of length 10 for 30 random seeds.

F.2   Deep RL experiments

We used DQN with the following hyperparameters:

Hyperparameter	Value
n-step update horizon	1
Discount (γ)	0.99
Adam epsilon	0.0003125
Minibatch size	64
Max replay buffer size	500000
Min replay buffer size	500
Learning rate	1e-4 linearly annealed to 1e-5
Epsilon greedy schedule	1 linearly annealed to 0.01 over first 10% of steps
Q-network update period	2
Target Q-network update period	50

Table 1: Shared DQN hyperparameters