Optimism and Delays in Episodic Reinforcement Learning

Benjamin Howson
Imperial College London

Ciara Pike-Burke
Imperial College London

Sarah Filippi
Imperial College London

Abstract

There are many algorithms for regret minimisation in episodic reinforcement learning. This problem is well-understood from a theoretical perspective, providing that the sequences of states, actions and rewards associated with each episode are available to the algorithm updating the policy immediately after every interaction with the environment. However, feedback is almost always delayed in practice. In this paper, we study the impact of delayed feedback in episodic reinforcement learning from a theoretical perspective and propose two general-purpose approaches to handling the delays. The first involves updating as soon as new information becomes available, whereas the second waits before using newly observed information to update the policy. For the class of optimistic algorithms and either approach, we show that the regret increases by an additive term involving the number of states, actions, episode length, the expected delay and an algorithm-dependent constant. We empirically investigate the impact of various delay distributions on the regret of optimistic algorithms to validate our theoretical results.

1 INTRODUCTION

Episodic Reinforcement Learning (RL) considers the problem of an agent learning how to act in an unknown environment to maximise its cumulative reward. The problem formulation is broad enough to capture the nature of sequential decision-making in many real-world scenarios as it permits complex dependencies between actions, rewards and future environmental states. Despite the complexity of the learning problem, there are many provably efficient algorithms for this problem setting (Jaksch et al., 2010; Filippi et al., 2010; Fruit et al., 2020; Azar et al., 2017; Dann et al., 2017).

These existing algorithms focus on the traditional model where one assumes that the algorithm updating the policy observes the sequence of states, actions and rewards at the end of every episode. Unfortunately, this immediate feedback assumption is unrealistic in almost all practical applications. In healthcare, for example, feedback relating to a patient on a particular treatment protocol is not observable to the policy maker until they return to the clinic at a scheduled time point in the future. In e-commerce, one observes a conversion at some unknown time long after a sequence of recommendations. Yet another example is wearable technology. Here, the heavy computation involved in policy updating must occur on a separate machine, forcing the communication of information, which naturally introduces a delay between the agent collecting feedback and the policy updater. In any of these scenarios, the algorithm must continue operating, despite lacking information from its past choices.

The above examples illustrate that delayed feedback is a fundamental challenge in real world reinforcement learning. Unfortunately, there is little theoretical understanding of the impact of delays in episodic reinforcement learning in the existing literature. We seek to fill this gap in the literature in this paper.

1.1 Related Work

Recently, the topic of delays has attracted a lot of attention in the bandit setting (Agarwal and Duchi, 2011; Dudik et al., 2011; Joulani et al., 2013; Mandel et al., 2015; Vernade et al., 2017; Pike-Burke et al., 2018; Zhou et al., 2019; Manegueu et al., 2020; Vernade et al., 2020). Here, the feedback is the reward associated with the chosen action in each round. Perhaps the most appealing approach in the multi-armed bandit setting is the queuing technique, which shows that the delays cause an additive penalty involving the expected delay for any base algorithm (Joulani et al., 2013; Mandel et al., 2015). The high-level idea is to build a meta-algorithm that creates a simulated non-delayed environment for any base algorithm designed for immediate feedback, such as UCB1 or KL-UCB. They achieve this by

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

introducing a mechanism that stores the rewards for each action in separate queues and having the base algorithm interact with these rather than the actual environment. Unfortunately, the queuing technique does not readily extend to the delayed feedback setting in RL, as forming the queues would require knowledge of the state and action seen in each step of an episode; this information is delayed in our setting.

Joulani et al. (2013) present another meta-algorithm for adversarial multi-armed bandits with delayed rewards that is trivial to adapt to our setting. They propose creating a new instance of the chosen base algorithm whenever there is no feedback, allowing one to bound the regret of each instance separately using standard techniques. More precisely, this involves maintaining $\tau_{\max} + 1$ versions of the algorithm, where $\tau \leq \tau_{\max}$ almost surely (Joulani et al., 2013). Thus, the regret of taking this approach is multiplicative, as the maximal delay scales the regret of the base algorithm.

Previous work in RL has considered constant delays in observing the current state in Markov Decision Processes (MDPs) (Katsikopoulos and Engelbrecht, 2003). More recent work considers delayed feedback in adversarial MDPs (Lancewicki et al., 2021). They developed an algorithm that computes stochastic policies based on policy optimisation. The regret of this algorithm depends on the sum of the delays, the number of states and the number of steps per episode. For stochastic MDPs, they state a regret bound of the form $H^{3/2}S\sqrt{AT} + H^2 S\tau_{\max}$, where H is the number of decisions the learner must make per episode, $T = KH$ is the total number of decisions made across all K episodes, S is the number of states in the environment, A is the number of actions and $\tau_k \leq \tau_{\max}$. However, the leading order term in their regret bound is loose for many base algorithms. Their approach also requires *a-priori* knowledge of the maximal delay to define a phase of explicit exploration; this quantity is often unknown in many practical applications. Further, the base algorithm accrues linear regret in this exploration phase, and the maximal delay can be prohibitively large. We propose two approaches that avoid such prior knowledge and can leverage new information in the early episodes much faster, leading to tighter algorithm-specific theoretical results and better empirical performance. In addition to the improved theoretical results, we relax the assumption that the delay distribution has a finite and known maximum, and instead only require that the delays have a finite expectation that we assume is unknown.

1.2 Contributions

The delayed feedback model studied in this paper poses several theoretical challenges that do not arise in the standard episodic reinforcement learning problem, such as delayed updates and disentangling the delays from the difficulty of the learning problem in the theoretical analysis.

We introduce two novel meta-algorithms to overcome these challenges, namely *active* and *lazy* updating. Both take any algorithm as input and transform it into an algorithm that can handle delayed feedback. Henceforth, we refer to the input algorithm as the *base algorithm*. Using these meta-algorithms, we obtain high probability regret bounds for any optimistic model-based base algorithm in the delayed feedback setting. For both active and lazy updating, the penalty for delayed feedback is an additive term involving the expected delay. Although they obtain similar theoretical results, active and lazy updating employ different algorithmic ideas to separate the delays from the learning problem in the theoretical analysis.

The active updating meta-algorithm uses the base algorithm to update the policy as soon as it observes feedback from the environment. Deriving theoretical guarantees for active updating involves tackling the delays head-on, as the delays force the policy to remain constant across numerous episodes. Consequently, the learner can repeatedly make sub-optimal decisions. To quantify the impact of delayed feedback, we introduce several techniques that carefully separate the difficulty of the learning problem from the delays.

The lazy meta-algorithm works slightly differently. Instead of updating immediately, it waits for the amount of feedback to surpass some threshold before updating the policy. One can control this threshold, and therefore the frequency of policy updates, through a hyperparameter α. By waiting to update, lazy creates a simulated non-delayed version of the environment for the input algorithm, allowing us to handle the delays separately from the difficulty of the learning problem.

2 PRELIMINARIES

We consider the task of learning to act optimally in an unknown episodic finite-horizon Markov Decision Process, EFH-MDP. An EFH-MDP is formalised as a quintuple: $M = (\mathcal{S}, \mathcal{A}, H, P, R)$. Here, $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, H is the horizon and gives the number of steps per episode, $P = \{P_h(\cdot|s,a)\}_{h,s,a}$ is the set of probability distributions over the next state and $R = \{R_h(s,a)\}_{h,s,a}$ is the set of reward functions. For conciseness, we assume that the reward function is known, deterministic and bounded between zero and one for all state-action-step triples.[1]

In the episodic reinforcement learning problem, the base algorithm interacts with an MDP in a sequence of episodes: $k = 1, 2, \ldots, K$. We denote the set of episodes by:

[1]The main challenge in model-based reinforcement learning lies in estimating the transition function. Thus, an extension to unknown bounded stochastic rewards is relatively straightforward.

$[K] = \{1, 2, \dots K\}$; a convention that we adopt for sets of integers. In this paper, we consider base algorithms that compute a deterministic policy $\pi_k : \mathcal{S} \times [H] \to \mathcal{A}$ at the start of each episode $k \in [K]$. It is known that in finite horizon stochastic MDPs, if an optimal policy exists, there is a deterministic optimal policy (Puterman, 1994). Once the base algorithm has computed a policy, an agent uses said policy to sample feedback from the environment by: selecting an action, $a_h^k = \pi_k(s_h^k, h)$; receiving a reward, $r_h^k = R_h(s_h^k, a_h^k)$; and transitioning to the next state, $s_{h+1}^k \sim P_h(\cdot|s_h^k, a_h^k)$; for each $h = 1, \cdots, H$. The feedback associated with the h-th step of the k-th episode is given by:

$$\mathcal{D}_h^k \coloneqq \{(s_h^k, a_h^k, r_h^k, s_{h+1}^k)\} \,. \tag{1}$$

We measure the quality of a policy, π, using the value function, which is the expected return at the end of the episode from the current step, given the current state:

$$V_h^\pi(s) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}^k \Big| s_{h'}^k = s \right] \,. \tag{2}$$

Further, we denote the optimal value function by: $V_h^*(s) = \max_\pi \{V_h^\pi(s)\}$, which gives the maximum expected return over deterministic policies $\forall (s,h) \in \mathcal{S} \times [H]$. When evaluating reinforcement learning algorithms, it is common to use regret:

$$\mathfrak{R}_K = \sum_{k=1}^{K} V_1^*\left(s_1^k\right) - V_1^{\pi_k}\left(s_1^k\right) \coloneqq \sum_{k=1}^{K} \Delta_1^k \,. \tag{3}$$

Throughout, $T = KH$ denotes the total number of steps. Domingues et al. (2020) show that the lower bound for the regret in the standard episodic reinforcement learning setting with stage-dependent transitions is: $\Omega(H\sqrt{SAT})$.

2.1 Regret Minimisation in Model-Based RL

Many provably efficient algorithms exist for learning in EFH-MDPs when feedback is immediate. In this paper, we focus on the large class of optimistic model-based reinforcement learning algorithms. These algorithms maintain estimators of the transition probabilities for each $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:

$$\hat{P}_{kh}\left(s'|s,a\right) = \frac{\sum_{i:i<k} \mathbb{1}\left\{s_{h+1}^i = s' \mid \left(s_h^i, a_h^i\right) = (s,a)\right\}}{N_{kh}\left(s,a\right)}$$

where

$$N_{kh}(s,a) = \max\left\{1, \sum_{i:i<k} \mathbb{1}\left\{\left(s_h^i = s, a_h^i = a\right)\right\}\right\}$$

is the total visitation count. There are two main ways of ensuring optimism using model-based algorithms. The first is the model-optimistic approach, which maintains a confidence set around $\hat{P}_{kh}$ that contains P_h with high probability (Jaksch et al., 2010; Filippi et al., 2010; Fruit et al., 2020). The second is the value-optimistic approach, which involves directly upper bounding the optimal value function with high probability by adding a bonus to the value function of a policy under the estimated transition density $\hat{P}_{kh}$ (Dann et al., 2017; Azar et al., 2017). Recent work has shown that all model-based optimistic algorithms have a value-optimistic representation, meaning they all compute a value function of the following form (Neu and Pike-Burke, 2020):

$$\tilde{V}_h^\pi = (H'+1) \wedge \left(R_h + \left\langle \hat{P}_{kh}, \tilde{V}_{h+1}^\pi \right\rangle + \beta_{kh}^+\right) \tag{4}$$

where $H' = H - h$ and

$$\begin{aligned}\beta_{kh}^+(s,a) &= H' \wedge \left(\frac{B_1}{\sqrt{N_{kh}(s,a)}} + \frac{B_2}{N_{kh}(s,a)}\right) \\ &= H' \wedge \beta_{kh}(s,a)\end{aligned} \tag{5}$$

is the exploration bonus and $x \wedge y = \min\{x, y\}$. Here, B_1 and B_2 are algorithm-dependent quantities which may depend on $S, A, H, \log(T)$ or the empirical variance of the optimistic value function. A suitably chosen exploration bonus ensures the computed value function is optimistic with high probability. For our theoretical results to hold, we require the following assumption on the base algorithm.

Assumption 1. *The exploration bonus upper bounds the estimation error with high probability. Mathematically:* $\beta_{kh}^+(s,a) \geq \langle (\hat{P}_{kh} - P_h)(\cdot|s,a), V_{h+1}^*(\cdot) \rangle$ *for all time-steps, with probability* $1-\delta$.

All value-optimistic algorithms explicitly use the estimation error to derive suitable bonuses. Further, model-optimistic algorithms compute bonuses satisfying this assumption implicitly (Neu and Pike-Burke, 2020). Therefore, Assumption 1 allows us to capture a wide range of model-based algorithms.

For our analysis, it will be helpful to define an algorithm-dependent variable C, which indicates whether the algorithm's bonuses satisfy the following inequality:

$$\beta_{kh}^+(s,a)) < \left\langle \left(\hat{P}_{kh} - P_h\right)(\cdot\,|s,a), \tilde{V}_{h+1}^{\pi_k}(\cdot) \right\rangle \tag{6}$$

for all s, a, h, k with probability $1-\delta$. Intuitively, $C = 1$ corresponds to a bonuses that sits somewhere between the estimation error and the difference between the expectation of the optimistic value function under the estimated and true transition function. Since these bonuses must sit within a specific (potentially narrow) interval, they are tighter. However, as we will see later, such bonuses come at the expense of lower-order terms. UBEV and UCBVI are algorithms where $C = 1$. Whereas UCRL2, UCRL2B, KL-UCRL and χ^2-UCRL are algorithms with $C = 0$.

3 DELAYED FEEDBACK

Under stochastic delays, the feedback from an episode does not return to the base algorithm immediately after the interaction. Instead, it returns at some unknown time in the future, $k + \tau_k$. Here, τ_k denotes the random delay between the agent playing the k^{th} episode and the base algorithm receiving the corresponding feedback. Throughout this paper, we make the following assumption about the delays:

Assumption 2. *The delays are positive, independent and identically distributed random variables with a finite expected value,* $\mathbb{E}[\tau_k] < \infty$.

The introduction of delays causes the feedback associated with an episode to return at some unknown time in the future, $k + \tau_k$. As a result, the base algorithm cannot update its policy using feedback from episode k at the start of episode $k + 1$. Instead, it can only use feedback it has observed, e.g. the feedback associated with episodes $i : i + \tau_i < k + 1$.

When working with delayed feedback in RL, it is helpful to introduce the observed and missing visitation counters:

$$N'_{kh}(s,a) = \sum_{i:i+\tau_i<k} \mathbb{1}\left\{\left(s_h^i, a_h^i\right) = (s,a)\right\} \tag{7}$$

$$N''_{kh}(s,a) = \sum_{i:i+\tau_i\geq k} \mathbb{1}\left\{\left(s_h^i, a_h^i\right) = (s,a)\right\}. \tag{8}$$

These are related to the total visitation counter by

$$N_{kh}(s,a) = N'_{kh}(s,a) + N''_{kh}(s,a)\ . \tag{9}$$

When the feedback is delayed, optimistic algorithms can only compute their bonuses and any required estimators using the observed visitation counter. The corresponding value functions are still optimistic, but they contract to the optimal value function more slowly since $N_{kh}(s,a) \geq N'_{kh}(s,a)$.

3.1 Bounding the Missing Episodes

In our analysis, it is helpful to bound the number of missing episodes to get an upper bound on the amount of information missing for each state-action-step. This is done in the following lemma.

Lemma 1. *Let* $S_k = \sum_{i=1}^{k-1} \mathbb{1}\{i + \tau_i \geq k\}$, *where* $\tau_1, \tau_2, \cdots \tau_{k-1} \sim f_\tau(\cdot)$ *are independent and identically distributed random variables with finite expected value. We define*

$$F_k^\tau = \left\{ S_k \geq \mathbb{E}[\tau] + \log\left(\frac{K\pi}{6\delta'}\right) + \sqrt{2\mathbb{E}[\tau]\log\left(\frac{K\pi}{6\delta'}\right)} \right\}$$

to be the failure event for a single k. *Then,* $\mathbb{P}(F_\tau) = \mathbb{P}(\cup_{k=1}^\infty F_k^\tau) \leq \delta'$.

Proof. Firstly, notice that S_k is a sum of Bernoulli random variables, meaning it is subgaussian. Therefore, one can apply Bernstein's inequality to obtain the following upper bound that holds with probability $1 - \delta'$:.

$$S_k \leq \mathbb{E}[S_k] + \frac{2}{3}\log\left(\frac{k\pi}{6\delta'}\right) + \sqrt{2\text{Var}\left(S_k\right)\log\left(\frac{k\pi}{6\delta'}\right)}$$

The remainder of the proof follows from noticing that $\mathbb{E}[S_k] \leq \sum_{i=0}^\infty \mathbb{P}(\tau > i)$, which is the tail probability function of the delay distribution and is equal to the expected delay. Similarly, one can show that $\text{Var}\left(S_k\right) \leq \mathbb{E}[S_k] \leq \mathbb{E}[\tau]$. Substituting these values into the above inequality gives the result. See Appendix A.1 for a full proof. □

A direct consequence of this lemma is an upper bound on the number of missing episodes $S_k \leq \psi_K^\tau$ for

$$\psi_K^\tau := \mathbb{E}\left[\tau\right] + \log\left(\frac{K\pi}{6\delta'}\right) + \sqrt{2\mathbb{E}\left[\tau\right]\log\left(\frac{K\pi}{6\delta'}\right)}$$

which holds for all $k \in [K]$ with probability $1 - \delta'$. Essentially, ψ_K^τ allows us to bound the amount of missing information in any given episode due to the delays.

4 META-ALGORITHMS FOR DELAYED FEEDBACK

Here, we describe two flexible approaches that allow any base algorithm to handle delayed feedback. Additionally, we prove regret guarantees for both procedures, providing the base algorithm satisfies Assumption 1. Regardless of the approach, we utilise the following regret decomposition for optimistic base algorithms that holds for both the delayed and non-delayed settings.

Lemma 2. *Under Assumption 1, with probability* $1 - 4\delta'$, *we can upper bound the regret by:*

$$\mathfrak{R}_K \leq 6\left(H + C\right)\sqrt{T\log\left(\frac{K\pi}{6\delta'}\right)} + 6\sum_{k=1}^K\sum_{h=1}^H \beta_{kh}^+\left(s_h^k, a_h^k\right) + 6\sum_{k=1}^K\sum_{h=1}^H \frac{3CH^2SL}{N'_{kh}\left(s_h^k, a_h^k\right)}$$

where $L = \log\left(S^2AH\pi^2/6\delta'\right)$ *and* C *indicates whether the bonuses of the algorithm satisfy Equation* (6).

Proof. See Appendix B.1. □

4.1 Active Updating

The first meta-algorithm we propose is *active updating*, which leverages new information by updating as soon as

Algorithm 1 Active Updating

Input. Base(N', M') (any base algorithm).
Initialise. $N' = \{N'_h(s,a) = 0\}_{h,s,a}$ and $M' = \{M'_h(s,a,s') = 0\}_{h,s,a}$ with

$$M'_h(s,a,s') \coloneqq \sum_{i: i+\tau_i < k} \mathbb{1}\{(s^i_h, a^i_h, s^i_{h+1}) = (s,a,s')\}$$

Compute policy: $\pi_1 = Base(N', M')$
for $k = 1$ **to** K **do**
  if $\exists\, i : k-2 < i + \tau_i \leq k-1$ **then**
    Update the counters: N' and M'.
    Update the policy: π_k = $Base(N', M')$
  else
    Reuse previous policy: $\pi_k = \pi_{k-1}$
  end if
  An agent samples an episode using policy π_k.
end for

it becomes available. The remainder of this subsection focuses on bounding the regret for model-based optimistic algorithms using active updating, whose pseudo-code is outlined in Algorithm 1.

Base(N', M') is the only input parameter for our algorithm and is the base algorithm. One could view it as a function that takes in the observed number of visits (N') and transitions (M'), among other algorithm-dependent hyperparameters, and returns a policy. For the class of optimistic algorithms, the additional hyperparameter is the confidence level, δ.

Theorem 1 (Active Updating)**.** *Under Assumption 1 and 2, with probability $1-\delta$, the regret of any model-based algorithm under delayed feedback:*

$$\mathfrak{R}_K \lesssim B\sqrt{HSAT} + \max\left\{B, B_2, CH^2S\right\} HSA\,\mathbb{E}\left[\tau\right]$$

where $\lesssim$ suppresses numeric constants, poly-log and lower order terms, and $B \geq B_1$ is a upper bound on the leading-order term in the numerator of the exploration bonus that is a function of H and S, and holds for all $(k,h) \in [K]\times[H]$.

Proof. From Lemma 2, it is clear that we must bound the summation of the bonuses to bound the regret. When there are no delays, one can utilise the fact that the visitation count for (s,a,h) at the start of episode $k+1$ increases by one if the agent observed (s,a,h) in the k-th episode to bound this term. However, this is no longer the case under delayed feedback. Therefore, we introduce the following lemma to bound the delay-dependent visitation counter.

Lemma 3. *Let $Z^p_T = \sum_{k=1}^K \sum_{h=1}^H 1/(N'_{kh}(s^k_h, a^k_h))^p$. Then,*

$$Z^p_T \leq \begin{cases} 4\sqrt{HSAT} + 3HSA\psi^\tau_K & \text{if } p = \frac{1}{2} \\ 2HSA\log(8T) + HSA\psi^\tau_K \log(16\psi^\tau_K) & \text{if } p = 1 \end{cases}$$

with probability $1-\delta'$.

Proof. To prove the claim, we relate the sum involving the observed visitation counters to a sum involving the total visitation counters. To do so, we artificially introduce it into the summation by multiplying by one:

$$Z^p_T = \sum_{k=1}^K \sum_{h=1}^H \left(\frac{N'_{kh}(s^k_h, a^k_h) + N''_{kh}(s^k_h, a^k_h)}{N'_{kh}(s^k_h, a^k_h) N_{kh}(s^k_h, a^k_h)}\right)^p$$
$$= \sum_{k=1}^K \sum_{h=1}^H \left(\frac{1}{N_{kh}(s^k_h, a^k_h)} + \frac{N''_{kh}(s^k_h, a^k_h)}{N'_{kh}(s^k_h, a^k_h) N_{kh}(s^k_h, a^k_h)}\right)^p$$

The term in the numerator of the first line is equivalent to the total visitation counter by the equivalence relation given in Equation (9). One can handle the first term using standard results from the immediate feedback setting. The remainder of the proof follows from carefully splitting the second term in the sum on the second line into two disjoint sets. Namely, we split the summation using two indicators: $\mathbb{1}\{N'_{kh}(s,a) \geq \psi^\tau_K\}$ and $\mathbb{1}\{N'_{kh}(s,a) < \psi^\tau_K\}$. After a little algebra, we find that we are able to apply results from the immediate feedback setting, which gives the final result. See Appendix A.2 for further details. □

For many algorithms, B_1 depends polynomially on quantities related to the environment, e.g. H and S. For such algorithms, a direct application of Lemma 3 is able to separate the expected delay from the total number of decisions. This is in line with the intuition that the impact of delays are negligible once we have a reasonable model of the environment. However, for algorithms such as for UCRL2B, χ^2-UCRL and UCBVI (Fruit et al., 2020; Neu and Pike-Burke, 2020; Azar et al., 2017):

$$B_1 = \widetilde{\mathcal{O}}\left(\sqrt{\mathbb{V}_{s'\sim\hat{P}_{kh}(\cdot\,|\,s,a)}\left(\tilde{V}^\pi_{h+1}(s')\right)}\right)$$

Typically, one uses an application of Cauchy-Schwarz to separate the terms involving the variance from those involving the counters, which gives:

$$\sum_{k,h} \sqrt{\frac{\mathrm{Var}_{s'\sim\hat{P}_h}(\tilde{V}_{h+1}(s'))}{N'_{kh}(s^k_h, a^k_h)}}$$
$$\leq \sqrt{\sum_{k,h} \mathrm{Var}_{s'\sim\hat{P}_h}(\tilde{V}_{h+1}(s')) \sum_{k,h} \frac{1}{N'_{kh}(s^k_h, a^k_h)}}$$

Lemma 3 shows that doing so would lead to the delays multiplying the leading order term, as the summation of the variances found underneath the square root is of order HT and multiplies the $HSA\psi^\tau_K$ that arises from bounding the summation of the observed visitation counter. Setting $B = H/2$ gives us an upper bound for these types of bonuses and avoids this multiplicative dependence.

Using a uniform upper bound on B_1 in conjunction with Lemma 3 allows us to handle the summation of the bonuses as follows:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{B_1}{\sqrt{N'_{kh}\left(s_h^k,a_h^k\right)}}+\frac{B_2+3CH^2SL}{N'_{kh}\left(s_h^k,a_h^k\right)}\\
&\leq\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{B}{\sqrt{N'_{kh}\left(s_h^k,a_h^k\right)}}+\frac{B_2+3CH^2SL}{N'_{kh}\left(s_h^k,a_h^k\right)}\\
&\leq 4B\sqrt{HSAT}+3BHSA\psi_K^\tau\\
&\quad+2\left(B_2+3CH^2SL\right)HSA\log(8T)\\
&\quad+\left(B_2+3CH^2SL\right)HSA\psi_K^\tau\log\left(16\psi_K^\tau\right)
\end{aligned}$$

Substituting the above upper bound of the terms in Lemma 2 and setting $\delta=5\delta'$ gives the stated result. □

Table 1 in Section 4.3 presents regret bounds for various optimistic algorithms using active updating under delayed feedback that fit into our framework. Further discussion of the results can be found in Section 4.3.

4.2 Lazy Updating

Instead of updating the policy via the base algorithm as soon as new feedback becomes observable, we now consider waiting. We name the meta-algorithm that employs this technique *lazy updating*. Algorithm 2 presents the pseudo-code for this meta-algorithm.

Algorithm 2 Lazy Updating

Input. Base$(N',M',\cdots)$ (any base algorithm) and α (activity parameter).
Initialise epoch: $j=1$ and $k_j=1$.
Initialise counters: $N'_{kh}(s,a)=M'_{kh}(s,a,s')=0$.
Compute policy: $\pi_{k_j}=Base(N'_{kh},M'_{kh})$
for $k=1$ **to** K **do**
  Update counters, e.g. Equation (7).
  if $\exists\,(s,a,h):N'_{kh}(s,a)\geq(1+1/\alpha)N'_{k_jh}(s,a)$ **then**
    Update epoch: $j=j+1$, $k_j=k$
    Update epoch counter: $N_{k_j}(s,a)=N'_{kh}(s,a)$
    Update the policy: $\pi_{k_j}=Base(N'_{k_jh},M'_{k_jh})$
  end if
  An agent samples an episode using policy π_{k_j}.
end for

Lazy updating works in batches of episodes which we call epochs and denote by $j=1,2,\cdots,J$. At the start of the j-th epoch, lazy updating uses the base algorithm to compute a policy using all the available information. The meta-algorithm uses this policy in every episode until the next epoch begins. Therefore, each epoch is just a set of episodes where the lazy updating algorithm uses the same policy.

A new epoch begins as soon as there is an (s,a,h) whose observed visitation counter reaches $1+1/\alpha$ times the observed visits at the start of the epoch, where $\alpha\in[1,\infty)$. Note that $\alpha=1$ corresponds to the well-known doubling trick from Jaksch et al. (2010), and $\alpha>1$ represents more frequent updating. Once the observed visitation counter triggers this condition, a new epoch begins, and the meta-algorithm uses the base algorithm to update the policy. Formally, we start epoch $j+1$ in episode k_{j+1}, which occurs when:

$$\begin{aligned}
k_{j+1}&=\arg\min_{k>k_j}\left\{N'_{kh}\geq\left(1+\frac{1}{\alpha}\right)N'_{k_jh}\right\}\\
&=\arg\min_{k>k_j}\left\{n_{k_j}^k\geq\frac{1}{\alpha}N'_{k_jh}\right\}
\end{aligned}\tag{10}$$

where

$$n_{kh}^l(s,a)=\sum_{i=k}^{l-1}\mathbb{1}\left\{(s_h^i=s,a_h^i=a),\,i+\tau_i\leq l\right\}\tag{11}$$

counts the observed number of visits between episodes k and l for $l>k$. Intuitively, this updating scheme forces the number of samples needed for any particular (s,a,h) to trigger an update to increase exponentially quickly, meaning that the total number of epochs should grow logarithmically in K. Lemma 4 confirms that this is indeed the case.

Lemma 4. *For $K\geq SA$ and $\alpha\geq1$, Algorithm 2 ensures that the number of epochs has the following upper bound:*

$$J\leq\frac{HSA\log\left(\frac{\alpha K}{SA}+1\right)}{\log(1+\frac{1}{\alpha})}$$

Proof. See Appendix A.3 for further details. □

In contrast to active updating, we will later see that the lazy updating scheme lets us bound the summation of the bonuses independently of the delays. This property means we can avoid upper bounding the numerator of the exploration bonus, B_1, and get tighter leading order terms in the regret bound of the chosen base algorithm. In the regret analysis, we will utilise the following extension of the classic result by Jaksch et al. (2010) that illustrates the delay-independence of the bonuses:

Lemma 5. *If $n_0,n_1,\cdots,n_J$ are an arbitrary sequence of real-valued numbers satisfying $n_0:=0$ and $0\leq n_j\leq\frac{1}{\alpha}N_{j-1}$ with $N_{j-1}=\max\{1,\sum_{i=0}^{j-1}n_i\}$ for all $j\leq J$, then*

$$\sum_{j=1}^{J}\frac{n_j}{N_{j-1}^p}\leq\begin{cases}\left(\sqrt{2}(1+\frac{1}{\alpha})+1\right)\sqrt{N_J} & \text{if } p=\frac{1}{2}\\ (1+\frac{1}{\alpha})+(1+\frac{1}{\alpha})\log\left(N_J\right) & \text{if } p=1\end{cases}$$

Proof. We prove the claim for each case using an inductive argument similar to Jaksch et al. (2010). See Appendix A.3. □

Using Lemmas 4 and 5, we can derive regret bounds for any optimistic base algorithm satisfying Assumption 1.

Theorem 2. *Let $K \geq SA$ and $\alpha \geq 1$. Under Assumption 1 and 2, with probability $1-\delta$, the regret of any model-based algorithm under delayed feedback is upper bounded by:*

$$\mathfrak{R}_K \lesssim \left(1 + \frac{1}{\alpha}\right) \hat{\mathfrak{R}}_K(Base) + \frac{H^2 SA \mathbb{E}[\tau]}{\log(1+\frac{1}{\alpha})}$$

where $\hat{\mathfrak{R}}_K(Base)$ is an upper bound on the regret of the chosen base algorithm under immediate feedback.

Proof. By optimism and utilising the fact that epochs are disjoint sets of episodes, with probability $1-\delta'$:

$$\begin{aligned}\mathfrak{R}_K \leq \tilde{\mathfrak{R}}_K &\coloneqq \sum_{k=1}^{K} \tilde{\Delta}_1^k\left(s_1^k\right) = \sum_{j=1}^{J} \sum_{k=k_j}^{k_{j+1}-1} \tilde{\Delta}_1^k\left(s_1^k\right) \\ &\leq HJ + \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \tilde{\Delta}_1^k\left(s_1^k\right)\end{aligned}$$

where the final inequality follows from separating the episodes where we update and bounding their contribution to the regret by HJ.

Handling the remaining summation in the regret bound requires a little more care, which we do by splitting the remaining sum into two sets; episodes with short and long delays. An episode has a short delay if it is played and observed in the same epoch, $\mathbb{1}\{k+\tau_k < k_{j+1}\}$. Otherwise, it has a long delay, $\mathbb{1}\{k+\tau_k \geq k_{j+1}\}$.

One can show that the regret of episodes with long delays has the following upper bound:

$$\sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \tilde{\Delta}_1^k\left(s_1^k\right) \mathbb{1}\{k+\tau_k \geq k_{j+1}\} \leq H \sum_{j=1}^{J} S_{k_{j+1}}$$

Aforementioned, $S_k \leq \psi_K^\tau$ for all $k \leq K$ with probability $1-\delta'$. Therefore, we can upper bound the regret of episodes with long delays by $HJ\psi_K^\tau$.

All that remains is bounding the regret of episodes with short delays. Applying Lemma 2 to these episodes and rearranging gives:[2]

$$\begin{aligned}&\sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \tilde{\Delta}_1^k\left(s_1^k\right) \mathbb{1}\{k+\tau_k < k_{j+1}\} \\ &\lesssim \sum_{s,a,h} \sum_{j=1}^{J} n_{k_{j+1}}^{k_{j+1}}(s,a) \beta_{kh}(s,a) \mathbb{1}\{k+\tau_k < k_{j+1}\}\end{aligned}$$

where we have omitted the state-action-step triples that caused the update from the summation. By construction, all the state-action-step triples satisfy the conditions of Lemma 5. Applying this result to the summation of the bonuses and combining the contributions of the other terms gives the result. See Appendix A.4 for a full proof of the claim. □

[2]Here, we have omitted lower order terms for brevity.

4.3 Discussion

Table 1 presents a selection of algorithms that fit into our framework and their accompanying theoretical guarantees when using the active and lazy updating meta-algorithms to handle delayed feedback. In particular, we see that acting in delayed environments causes an additive increase in regret for almost all combinations of optimistic base algorithms and meta-algorithms considered. This result mirrors what is seen in the bandit setting where algorithms incur an additive regret penalty involving $\mathbb{E}[\tau]$ (Joulani et al., 2013).

For active updating and some base algorithms, we found that the additive delay dependence comes at the price of a penalty to the leading order term in the regret bound. Namely, an extra $\sqrt{H}$. This extra penalty multiplying the leading order term is a feature of the theoretical analysis. Another important factor influencing the impact of the delays when using active updating is the parameter C. The penalty for delayed feedback is higher when $C = 1$. The worsened delay dependence for these algorithms is due to the introduction of lower-order terms in the probabilistic analysis under immediate feedback, which allows for tighter bonuses. Unfortunately, these lower-order terms become dependent on the delays in our setting and thus lead to a worse delay dependence.

To rectify the undesirable penalty to the leading order terms and the dependence on C, we developed an alternative approach called lazy updating, which achieves the same additive delay dependence for all algorithms that fit into our framework with only a logarithmic penalty to the leading order term in the regret bound of the base algorithm under immediate feedback. This approach works by introducing an additional hyperparameter that controls how frequently the base algorithm updates its policy. We denote this hyperparameter by α and name it the activity parameter. Theorem 2 indicates that there is a trade-off when selecting α. On the one hand, we would like to choose a large value of α to minimise the penalty to the leading order term, which is arises from the slower updating. On the other hand, the penalty introduced by the delays is a strictly increasing function of α, making large values undesirable. As $\alpha \to \infty$, lazy updating tends to active updating; at this limiting value, lazy updating will update as soon as it receives new feedback, just like active updating. Thus, the empirical performance of lazy updating should get closer to active updating as α increases. In Section 5, we demonstrate that this is the case and show that it is pos-

Table 1: Example Delayed Feedback Regret Bounds

Base Algorithm	C	$\hat{\mathfrak{R}}_K(\text{Base})$	**Active Updating**	**Lazy Updating**
UBEV (Dann et al., 2017)	1	$H^{3/2}\sqrt{SAT}$	$\hat{\mathfrak{R}}_K(\text{Base}) + H^3S^2A\mathbb{E}[\tau]$	$(1+\frac{1}{\alpha})\hat{\mathfrak{R}}_K(Base) + \frac{H^2SA\mathbb{E}[\tau]}{\log(1+\frac{1}{\alpha})}$
UCBVI-CH (Azar et al., 2017)	1	$H^{3/2}\sqrt{SAT}$	$\hat{\mathfrak{R}}_K(\text{Base}) + H^3S^2A\mathbb{E}[\tau]$	$(1+\frac{1}{\alpha})\hat{\mathfrak{R}}_K(Base) + \frac{H^2SA\mathbb{E}[\tau]}{\log(1+\frac{1}{\alpha})}$
UCRL2 (Jaksch et al., 2010)	0	$H^{3/2}S\sqrt{AT}$	$\hat{\mathfrak{R}}_K(\text{Base}) + H^2S^{3/2}A\mathbb{E}[\tau]$	$(1+\frac{1}{\alpha})\hat{\mathfrak{R}}_K(Base) + \frac{H^2SA\mathbb{E}[\tau]}{\log(1+\frac{1}{\alpha})}$
KL-UCRL (Filippi et al., 2010)	0	$H^{3/2}S\sqrt{AT}$	$\hat{\mathfrak{R}}_K(\text{Base}) + H^2S^{3/2}A\mathbb{E}[\tau]$	$(1+\frac{1}{\alpha})\hat{\mathfrak{R}}_K(Base) + \frac{H^2SA\mathbb{E}[\tau]}{\log(1+\frac{1}{\alpha})}$
UCRL2B (Fruit et al., 2020)	0	$H\sqrt{S\Gamma AT}$	$\sqrt{H}\hat{\mathfrak{R}}_K(\text{Base}) + H^2S^2A\mathbb{E}[\tau]$	$(1+\frac{1}{\alpha})\hat{\mathfrak{R}}_K(Base) + \frac{H^2SA\mathbb{E}[\tau]}{\log(1+\frac{1}{\alpha})}$
χ^2-UCRL (Neu and Pike-Burke, 2020)	0	$HS\sqrt{AT}$	$\sqrt{H}\hat{\mathfrak{R}}_K(\text{Base}) + H^2S^2A\mathbb{E}[\tau]$	$(1+\frac{1}{\alpha})\hat{\mathfrak{R}}_K(Base) + \frac{H^2SA\mathbb{E}[\tau]}{\log(1+\frac{1}{\alpha})}$
UCBVI-BF (Azar et al., 2017)	1	$H\sqrt{SAT}$	$\sqrt{H}\hat{\mathfrak{R}}_K(\text{Base}) + H^3S^2A\mathbb{E}[\tau]$	$(1+\frac{1}{\alpha})\hat{\mathfrak{R}}_K(Base) + \frac{H^2SA\mathbb{E}[\tau]}{\log(1+\frac{1}{\alpha})}$

sible to get most of the benefits of active updating with a relatively modest value of α, which has better worst-case regret bounds in the delayed feedback setting.

Comparatively, our work significantly improves the regret bounds for many algorithms in the delayed feedback setting. Lancewicki et al. (2021) presents regret bounds for stochastic MDPs of the form $H^{3/2}S\sqrt{AT} + H^2S\tau_{\max}$ for all optimistic algorithms. Except for UCRL2 and KL-UCRL, the leading order term in their regret bound is loose in either H, S or both. Conversely, the leading order terms in our regret bounds are tight for all algorithms when utilising lazy updating and are only loose by a factor of $\sqrt{H}$ for a few algorithms when utilising active updating. Furthermore, $\mathbb{E}[\tau] \ll \tau_{\max}$ in almost all scenarios. As a result, our regret bounds have a tighter delay dependence. Our algorithms also remove the need for *a-priori* knowledge of the maximal delay.

The setting of delayed feedback also generalises the case where only the rewards are delayed. Thus, our theoretical results also hold for this setting if we directly apply active or lazy updating. However, one could do better in this case by realising that it is only the delays impacting the rewards, meaning it is only necessary to apply the meta-algorithms to the estimation of the rewards. We expect the additive penalty to be $HSA\mathbb{E}[\tau]$. Indeed, the improved delay-dependence is due to the fact that learning the expected reward function is an easier task than learning the transitions. We prove that this is indeed the case for UCRL2 algorithm of Jaksch et al. (2010) in Appendix B.2.

5 EXPERIMENTAL RESULTS

In this section, we investigate the impact of delayed feedback on the regret of active and lazy updating in the chain environment of Osband and Van Roy (2017). Briefly, this environment consists of a sequence of S states arranged side-by-side. The learner starts in the left-most state and has to decide between $A = 2$ actions, head left or right. Each episode consists of $H = S$ decisions and the only state with a reward is the right-most state. Thus, the optimal policy is to head right at every step. Heading left is always successful. However, heading right is successful with probability $1 - 1/S$. If unsuccessful, the learner moves one state to the left. Notably, any inefficient exploration strategy will take at least 2^S episodes to learn the optimal policy (Osband and Van Roy, 2017).

We consider chains with $H = S \in \{5, 10, 20, 30\}$ and use UCBVI-BF as the base algorithm in all of our experiments as it has the best regret guarantees under immediate feedback. For our lazy updating approach, we selected several values for the activity hyperparameter, $\alpha \in \{1, 10, 100\}$. In all our experiments, we set the confidence parameter of the base algorithm so that the regret bounds hold with probability 0.95. Additionally, we compare our meta-algorithms to the explicit exploration procedure proposed by Lancewicki et al. (2021). Their procedure requires prior knowledge of the maximum delay, which we provide by generating all the delays before the first episode and taking the maximum. In practice, the maximum delay is often unknown and possibly infinite, making this approach infeasible.

Our experiments consider Constant, Geometric, Poisson and Uniform delays. For each of these distributions, we consider the following expected delays: $\mathbb{E}[\tau] \in \{0, 100, 200, 300, 400, 500\}$.[3] All results are averaged over 30 independent runs and the shaded regions in all the figures contain 95% of our empirical results.

Figure 1 displays the results for our experiments in the chain environment with $S = 30$ and $\mathbb{E}[\tau] = 100$. The results for the other chain lengths and expected values are in Appendix C. Empirically, active updating achieves the best performance of all three meta-algorithms. However, our experimental results suggest that it is possible to get near identical performance with lazy updating by setting α to be a large enough constant. Both active and lazy updating offer superior performance to the explicit exploration approach of Lancewicki et al. (2021) in all of our exper-

[3]For the uniformly distributed delays, we set the lower and upper limits to 0 and $2\mathbb{E}[\tau]$, respectively.

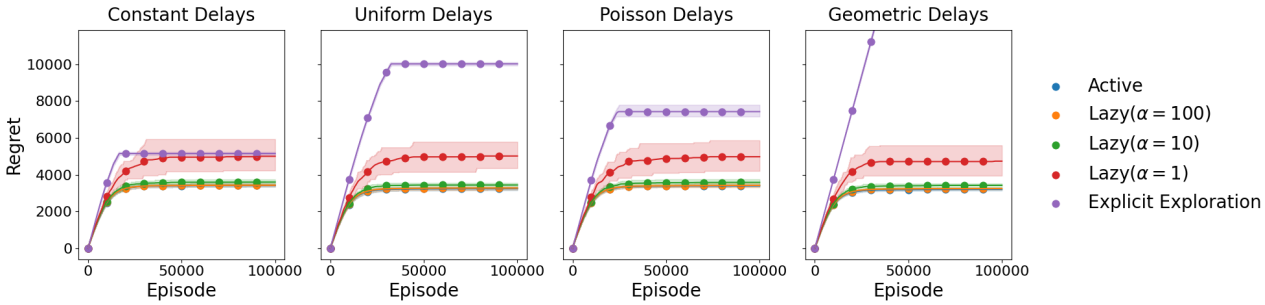

Figure 1: Cumulative Regret ($S = 30, \mathbb{E}[\tau] = 100$).

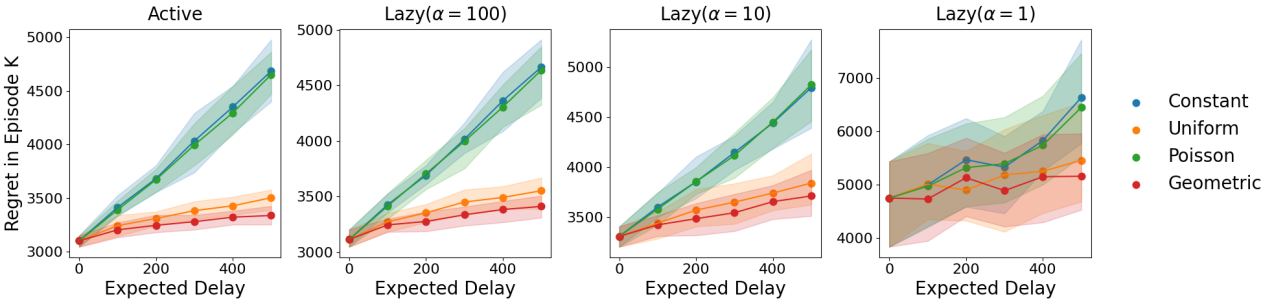

Figure 2: Delay Dependence ($S = 30$).

iments, despite their meta-algorithm having prior knowledge of the delays. In some cases, our meta-algorithms have converged to the optimal policy before the explicit exploration procedure finishes; e.g. see Appendix C.

Next, we turn to considering the impact of different delay distributions on the regret of our meta-algorithms. Empirically, Figure 2 shows that the regret penalty of delays at the end of the final episode is linear in the expected delay for active updating and lazy updating, as our theory predicts. For lazy updating, the gradient of this linear relationship decreases with α, which is to be expected based on the $\log(1 + 1/\alpha)$ term in the denominator of the delay-dependent terms in our regret bounds. Interestingly, lazy updating with $\alpha = 1$ is the most robust to the delay distribution. We believe that this is due to forcing the base algorithm to wait for long periods of time between updates. Intuitively, if the epochs are long enough, most information within an epoch will be received before an update, leading to little loss of information. Investigating this further is an interesting avenue for future work.

6 CONCLUSION

In this paper, we provide two generic meta-algorithms that can extend any episodic reinforcement learning base algorithm to the setting of delayed feedback. Under mild assumptions on the algorithm and the delays, we show that both maintain the sub-linear theoretical guarantees of the chosen base algorithm and provide good empirical performance, regardless of the delay distribution. These first positive results for stochastically delayed feedback in episodic reinforcement learning prove that the penalty for delays is an additive term involving the expected delay that is independent of the number of episodes. This additive penalty matches what is seen in the multi-armed bandit setting, despite the additional complexities of the reinforcement learning problem.

Our framework is broad enough to cover the theoretically successful class of optimistic model-based algorithms, and many existing algorithms fit into our framework. However, we believe that both updating procedures could be used for a wider class of base algorithms. For example, model-free optimistic algorithms and posterior sampling (Jin et al., 2018; Osband and Van Roy, 2017). Extending our analyses to cover these algorithms is left to future work.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful feedback that helped greatly improved the clarity and quality of the manuscript.

BH is funded by EPSRC through the Modern Statistics and Statistical Machine Learning CDT. Grant number EP/S023151/1.

References

Alekh Agarwal and John C Duchi. Distributed Delayed Stochastic Optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 263–272. PMLR, 2017.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 5717–5727. Curran Associates Inc., 2017.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic Reinforcement Learning in Finite MDPs: Minimax Lower Bounds Revisited. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, 2020.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient Optimal Learning for Contextual Bandits. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, page 169–178. AUAI Press, 2011.

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in Reinforcement Learning and Kullback-Leibler Divergence. In *48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122, 2010.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Improved Analysis of UCRL2 with Empirical Bernstein Inequality, 2020. URL https://arxiv.org/abs/2007.05456.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-Optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11:1563–1600, August 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Pooria Joulani, András György, and Csaba Szepesvári. Online Learning under Delayed Feedback. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, volume 28, page 1453–1461. JMLR.org, 2013.

K.V. Katsikopoulos and S.E. Engelbrecht. Markov decision processes with delays and asynchronous cost collection. *IEEE Transactions on Automatic Control*, 48 (4):568–574, 2003.

Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. Learning Adversarial Markov Decision Processes with Delayed Feedback, 2021. URL https://arxiv.org/abs/2012.14843.

Friedrich Liese and Igor Vajda. On Divergences and Informations in Statistics and Information Theory. *IEEE Transactions on Information Theory*, 52:4394–4412, 2006.

Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. The Queue Method: Handling Delay, Heuristics, Prior Data, and Evaluation in Bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015. doi: 10.1609/aaai.v29i1.9604. URL https://ojs.aaai.org/index.php/AAAI/article/view/9604.

Anne Gael Manegueu, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic Bandits with Arm-Dependent Delays. In *Proceedings of the 37th International Conference on International Conference on Machine Learning - Volume 28*. JMLR.org, 2020.

Gergely Neu and Ciara Pike-Burke. A Unifying View of Optimism in Episodic Reinforcement Learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2020.

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*, ICML'17, page 2701–2710. JMLR.org, 2017.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4105–4113. PMLR, 10–15 Jul 2018.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic Bandit Models for Delayed Conversions. In *In Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckner. Linear bandits with stochastic delayed feedback, 2020. URL https://arxiv.org/abs/1807.02089.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in Generalized Linear Contextual Bandits with Stochastic Delays. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/56cb94cb34617aeadff1e79b53f38354-Paper.pdf.

A MISSING PROOFS

A.1 Bounding the Missing Episodes

An important aspect in our proofs is to bound the amount of missing information. Since we see only one state-action pair per step of an episode, an upper bound on the missing visitation counter is simply the number of missing episodes. Lemma 1 bounds the number of missing episodes with high probability and only requires the delays have a finite expected value.

Lemma 1. *Let $S_k = \sum_{i=1}^{k-1} \mathbb{1}\{i + \tau_i \geq k\}$, where $\tau_1, \tau_2, \cdots \tau_{k-1} \sim f_\tau(\cdot)$ are independent and identically distributed random variables with finite expected value. We define*

$$F_k^\tau = \left\{ S_k \geq \mathbb{E}\left[\tau\right] + \log\left(\frac{K\pi}{6\delta'}\right) + \sqrt{2\mathbb{E}\left[\tau\right]\log\left(\frac{K\pi}{6\delta'}\right)} \right\}$$

to be the failure event for a single k. Then, $\mathbb{P}(F_\tau) = \mathbb{P}(\cup_{k=1}^{\infty} F_k^\tau) \leq \delta'$.

Proof. By definition, the summation involves a sequence of independent indicator random variables. Considering its expectation reveals that:

$$\begin{aligned}
\mathbb{E}\left[S_k\right] &= \sum_{i=1}^{k-1} \mathbb{E}\left[\mathbb{1}\left\{i + \tau_i \geq k\right\}\right] = \sum_{i=1}^{k-1} \mathbb{P}\left[\mathbb{1}\left\{i + \tau_i \geq k\right\}\right] = \sum_{i=1}^{k-1} \mathbb{P}\left[\,\tau_{k-i} > i\,\right] = \sum_{i=0}^{k-2} \mathbb{P}\left[\,\tau_{k-i+1} > i\,\right] \\
&\leq \sum_{i=0}^{\infty} \mathbb{P}\left[\,\tau > i\,\right] = \sum_{i=0}^{\infty} \sum_{j=i+1}^{\infty} \mathbb{P}\left[\,\tau = j\,\right] = \sum_{j=1}^{\infty} \sum_{i=0}^{j-1} \mathbb{P}\left[\,\tau = j\,\right] = \sum_{j=1}^{\infty} j\,\mathbb{P}\left[\,\tau = j\,\right] \\
&= \mathbb{E}\left[\tau\right].
\end{aligned}$$

Next, looking at its variance reveals that:

$$\begin{aligned}
\mathrm{Var}\left(S_k\right) &= \sum_{i=1}^{k-1} \mathrm{Var}\left(\mathbb{1}\left\{i + \tau_i \geq k\right\}\right) = \sum_{i=1}^{k-1} \mathbb{E}\left[\left(\mathbb{1}\left\{i + \tau_i \geq k\right\} - \mathbb{E}\left[\mathbb{1}\left\{i + \tau_i \geq k\right\}\right]\right)^2\right] \\
&\leq \sum_{i=1}^{k-1} \mathbb{E}\left[\mathbb{1}\left\{i + \tau_i \geq k\right\}^2\right] = \sum_{i=1}^{k-1} \mathbb{E}\left[\mathbb{1}\left\{i + \tau_i \geq k\right\}\right] = \mathbb{E}\left[S_k\right] \\
&\leq \mathbb{E}\left[\tau\right]
\end{aligned}$$

By Bernstein's inequality, we have that:

$$\mathbb{P}\left(S_k - \mathbb{E}\left[S_k\right] \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{\mathrm{Var}\left(S_k\right) + \frac{\epsilon}{3}}\right) = \frac{6\delta'}{\left(k\pi\right)^2}$$

Rearranging the above reveals that:

$$\epsilon \leq \frac{1}{3}\log\left(\frac{\left(k\pi\right)^2}{6\delta'}\right) + \sqrt{\mathrm{Var}\left(S_k\right)\log\left(\frac{\left(k\pi\right)^2}{6\delta'}\right)} \leq \frac{2}{3}\log\left(\frac{k\pi}{6\delta'}\right) + \sqrt{2\mathbb{E}\left[\tau\right]\log\left(\frac{k\pi}{6\delta'}\right)}$$

Since $k \leq K$, we have that:

$$\mathbb{P}\left(F_k^\tau\right) = \mathbb{P}\left(S_k - \mathbb{E}\left[\tau\right] \geq \frac{2}{3}\log\left(\frac{K\pi}{6\delta'}\right) + \sqrt{2\mathbb{E}\left[\tau\right]\log\left(\frac{K\pi}{6\delta'}\right)}\right) \leq \frac{6\delta'}{(k\pi)^2}$$

By Boole's inequality, we have that:

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} F_k^\tau\right) \leq \sum_{k=1}^{\infty} \mathbb{P}\left(F_k^\tau\right) = \frac{6\delta'}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{k^2} = \delta'$$

as required. □

A.2 Missing Proofs for Active Updating

Lemma 2 (the regret decomposition) and Equation (5) (the form of the exploration bonuses) reveal that the summation of the counters is an important quantity in determining the regret of an optimistic algorithm. Whenever $\tau_k = 0$ for all $k \leq K$, e.g. immediate feedback, we can use standard results that utilise the fact the counters increase by one between successive plays of a state-action pair at a given step.

Lemma 6. *Let $Z_n^p = \sum_{n=0}^{N} 1/(1 \vee n)^p$. Then, Z_n^p has the following upper bound:*

$$Z_n^p \leq \begin{cases} 2\sqrt{N} & \text{if } p \in \frac{1}{2} \\ \log(8N) & \text{if } p = 1 \end{cases}$$

for $p = 1/2$ and $p = 1$.

Proof. Removing the first two terms from the summation and upper bounding the remaining terms by an integral gives:

$$\begin{aligned} Z_n^p &= 2 + \sum_{n=2}^{N} \frac{1}{n^p} \leq 2 + \int_1^N \frac{1}{n^p} dn \leq 2 + \begin{cases} 2\sqrt{N} - 2 & \text{if } p \in \frac{1}{2} \\ \log(N) & \text{if } p = 1 \end{cases} \\ &\leq \begin{cases} 2\sqrt{N} & \text{if } p \in \frac{1}{2} \\ \log(8N) & \text{if } p = 1 \end{cases} \end{aligned}$$

as required. □

When τ_k is random, the observed visitation counter need not increase by one between successive plays of the same state-action-step. Instead, the counter only increases by one (or more in some cases) after a random number of episodes. In the worst-case scenario, the counter will remain constant between playing and observing the feedback associated with a specific state-action-step. Thus, the standard techniques no longer apply, and we must find another way to bound the summation of counters than can remain unchanged for numerous episodes due to the delays. We do this by relating the summation involving the observed visitation counter to one involving the total visitation counter, thereby splitting the terms affected by the delays from those that are not.

Lemma 3. *Let $Z_T^p = \sum_{k=1}^{K} \sum_{h=1}^{H} 1/(N'_{kh}(s_h^k, a_h^k))^p$. Then,*

$$Z_T^p \leq \begin{cases} 4\sqrt{HSAT} + 3HSA\psi_K^\tau & \text{if } p = \frac{1}{2} \\ 2HSA\log(8T) + HSA\psi_K^\tau \log(16\psi_K^\tau) & \text{if } p = 1 \end{cases}$$

with probability $1 - \delta'$.

Proof. Unless otherwise stated, we let: $N_{kh}(s,a) = 1 \vee N_{kh}(s,a)$ and $N'_{kh}(s,a) = 1 \vee N'_{kh}(s,a)$ for notational convenience. First, we use the relationships between the observed, missing and total visitation counters to split the summation into two parts. To do so, in a similar manner to Lancewicki et al. (2021), we start by artificially introducing the total visitation counter:

$$Z_T^p = \sum_{k=1}^{K} \sum_{h=1}^{H} \left(\frac{1}{N'_{kh}\left(s_h^k, a_h^k\right)} \right)^p = \sum_{k,h} \left(\frac{1}{N_{kh}\left(s_h^k, a_h^k\right)} \right)^p \left(\frac{N_{kh}\left(s_h^k, a_h^k\right)}{N'_{kh}\left(s_h^k, a_h^k\right)} \right)^p$$

From Equation (9), $N_{kh}(s,a) = N'_{kh}(s,a) + N''_{kh}(s,a)$, for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Consequently,

$$Z_T^p \leq \underbrace{\sum_{k,h} \left(\frac{1}{N_{kh}\left(s_h^k, a_h^k\right)} \right)^p}_{(i)} + \underbrace{\sum_{k,h} \left(\frac{1}{N_{kh}\left(s_h^k, a_h^k\right)} \frac{N''_{kh}\left(s_h^k, a_h^k\right)}{N'_{kh}\left(s_h^k, a_h^k\right)} \right)^p}_{(ii)},$$

since $(1+x)^p \leq 1 + x^p$ for $p = 1/2$ and $p = 1$ and any $x > 0$. Term (i) is the summation of the total visitation counter. Thus, Lemma 6 applies.

Bounding (ii) requires more care, as it involves the observed and missing visitation counters. Recall that the algorithm plays one state-action pair at each step in every episode. Thus, the missing visitation counter is upper bounded by the number of missing episodes: $N''_{kh}(s,a) \leq S_k$. Lemma 1 bounds the number of missing episodes: with probability $1-\delta'$, $S_k \leq \psi_K^\tau$ across all $k \in \mathbb{Z}^+$. Splitting (ii) using the observed visitation counts and the upper bound on S_k gives:

$$\begin{aligned}(ii) &\leq \sum_{k,h}\left(\frac{\mathbb{1}\left\{N'_{kh}\left(s_h^k,a_h^k\right)\geq \psi_K^\tau\right\}\psi_K^\tau}{N_{kh}\left(s_h^k,a_h^k\right)N'_{kh}\left(s_h^k,a_h^k\right)}\right)^p + \sum_{k,h}\left(\frac{\mathbb{1}\left\{N'_{kh}\left(s_h^k,a_h^k\right)\leq \psi_K^\tau\right\}\psi_K^\tau}{N_{kh}\left(s_h^k,a_h^k\right)N'_{kh}\left(s_h^k,a_h^k\right)}\right)^p \\ &\leq \underbrace{\sum_{k,h}\left(\frac{\mathbb{1}\left\{N'_{kh}\left(s_h^k,a_h^k\right)\geq \psi_K^\tau\right\}}{N_{kh}\left(s_h^k,a_h^k\right)}\right)^p}_{(ii.a)} + \underbrace{\sum_{k,h}\left(\frac{\mathbb{1}\left\{N'_{kh}\left(s_h^k,a_h^k\right)\leq \psi_K^\tau\right\}\psi_K^\tau}{N_{kh}\left(s_h^k,a_h^k\right)N'_{kh}\left(s_h^k,a_h^k\right)}\right)^p}_{(ii.b)}\end{aligned}$$

The last inequality follows since for the first sum, $N'_{kh}(s,a) \geq \psi_K^\tau$.

Clearly, $(ii.a) \leq (i)$, as it is a summation over a subset of all the episodes. Using (9), it is possible to rewrite the indicator in the remaining term as: $\mathbb{1}\{N_{kh}(s,a) - N''_{kh}(s,a) \leq \psi_K^\tau\}$, for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$. Further, $N''_{kh}(s,a) \leq \psi_K^\tau$ and $N'_{kh}(s,a) \geq 1$. Therefore,

$$\begin{aligned}(ii.b) &\leq (\psi_K^\tau)^p \sum_{k,h}\left(\frac{\mathbb{1}\left\{N_{kh}\left(s_h^k,a_h^k\right)\leq 2\psi_K^\tau\right\}}{N_{kh}\left(s_h^k,a_h^k\right)}\right)^p \\ &\leq (\psi_K^\tau)^p \sum_{s,a,h}\sum_{n=0}^{2\psi_K^\tau}\frac{1}{(1\vee n)^p}\end{aligned}$$

Lemma 6 gives an upper bound of $\sum_{n=0}^{N} 1/(1\vee n)^p$. Summing this upper bound over all state-action-step triples gives:

$$(ii.b) \leq \begin{cases} 3HSA\psi_K^\tau & \text{if } p=\frac{1}{2} \\ HSA\psi_K^\tau \log\left(16\psi_K^\tau\right) & \text{if } p=1\end{cases}$$

Therefore:

$$\begin{aligned}Z_T^p &\leq 2A + B.2 \\ &\leq \begin{cases} 4\sqrt{HSAT} + 3HSA\psi_K^\tau & \text{if } p=\frac{1}{2} \\ HSA\left(2\log\left(8T\right) + \psi_K^\tau\log\left(16\psi_K^\tau\right)\right) & \text{if } p=1\end{cases}\end{aligned}$$

as required. □

A.3 Missing Proofs for Lazy Updating

When using active updating, we prove that the bound on the counts depends on the delay. However, we can mitigate this delay-dependence by taking a slower approach to updating, providing that the number of epochs is bounded and the counts between epochs satisfy certain constraints outlined in Section 4.2.

Lemma 4. *For $K \geq SA$ and $\alpha \geq 1$, Algorithm 2 ensures that the number of epochs has the following upper bound:*

$$J \leq \frac{HSA\log\left(\frac{\alpha K}{SA}+1\right)}{\log(1+\frac{1}{\alpha})}$$

Proof. In this proof, we extend arguments from the standard doubling trick of Jaksch et al. (2010) so that the learner can update more frequently. Firstly, we recall the definition of the observed visitation counter:[4]

$$N'_k(s,a,h) = \sum_{i=1}^{k-1}\mathbb{1}\left\{\left(s_h^i,a_h^i\right)=(s,a), i+\tau_i < k\right\}$$

[4]We move the subscript denoting the step into the bracket for notational convenience

and the updating rule for $j \geq 1$:

$$k_{j+1} = \underset{k>k_j}{\arg\min} \left\{ \exists s, a, h : N'_k(s,a,h) \geq \left(1 + \frac{1}{\alpha}\right) N'_{k_j}(s,a,h) \right\}$$

Now, we define a counter that counts the observed number of visits between two episodes:

$$n_k^l(s,a,h) = \sum_{i=1}^{l-1} \mathbb{1}\left\{ \left(s_h^i, a_h^i\right) = (s,a), k \leq i + \tau_i < l \right\}$$

Direct computation allows us to relate the observed visitation counter at the start of the $(j+1)$-th epoch to the sum of the observed visitation counts within each of the previous epochs:

$$\begin{aligned}
N'_{k_{j+1}}(s,a,h) &= \sum_{i=1}^{k_{j+1}-1} \mathbb{1}\left\{ \left(s_h^i, a_h^i\right) = (s,a), i + \tau_i < k \right\} \\
&= \sum_{l=1}^{j} \sum_{i=k_l}^{k_{l+1}-1} \mathbb{1}\left\{ \left(s_h^i, a_h^i\right) = (s,a), i + \tau_i < k \right\} \\
&= \sum_{l=1}^{j} \sum_{i=1}^{k_{j+1}} \mathbb{1}\left\{ \left(s_h^i, a_h^i\right) = (s,a), k_l \leq i + \tau_i < k \right\} \\
&= \sum_{l=1}^{j} n_{k_l}^{k_{l+1}}(s,a,h)
\end{aligned}$$

where the second equality follows from the fact that an epoch is a disjoint set of episodes and the final equality follows from the definition of the between episodes visitation counter. From the above, it is easy to see that

$$N'_{k_{j+1}}(s,a,h) = n_{k_j}^{k_{j+1}}(s,a,h) + \sum_{l=1}^{j-1} n_{k_l}^{k_{l+1}}(s,a,h) = n_{k_j}^{k_{j+1}}(s,a,h) + N'_{k_j}(s,a,h)$$

Thus, we can re-write the updating rule using the within episode counter as:

$$\begin{aligned}
k_{j+1} &= \underset{k>k_j}{\arg\min} \left\{ \exists s, a, h : N'_k(s,a,h) \geq \left(1 + \frac{1}{\alpha}\right) N'_{k_j}(s,a,h) \right\} \\
&= \underset{k>k_j}{\arg\min} \left\{ \exists s, a, h : N'_k(s,a,h) - N'_{k_j}(s,a,h) \geq \frac{1}{\alpha} N'_{k_j}(s,a,h) \right\} \\
&= \underset{k>k_j}{\arg\min} \left\{ \exists s, a, h : n_{k_j}^{k_{j+1}}(s,a,h) \geq \frac{1}{\alpha} N'_{k_j}(s,a,h) \right\}
\end{aligned}$$

providing that we have seen the state-action-step at least once.[5] Therefore, at the end of each epoch there is a state-action-step with $n_{k_j}^{k_{j+1}}(s,a,h) \geq N'_{k_j}(s,a,h)/\alpha$.

Suppose $N'_{(K+1)h}(s,a) > 0$ for a fixed $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Define $J(s,a,h)$ as the number of epochs with $n_{k_j}^{k_{j+1}}(s,a,h) \geq N'_{k_j h}(s,a)/\alpha$. Or, equivalently, it is the number of epochs with $N_{k_{j+1}}(s,a,h) \geq (1+1/\alpha) N'_{k_j}(s,a,h)$.

[5] We handle the case for the epochs where the observed visitation count is zero later on in the proof.

Then,

$$\begin{aligned}
N'_{K+1}(s,a,h) &= \sum_{j=1}^{J} n_{k_j}^{k_{j+1}}(s,a,h) \\
&\geq 1 + \sum_{j: n_{k_j}^{k_{j+1}}(s,a,h) \geq N'_{k_j}(s,a,h)/\alpha} n_{k_j}^{k_{j+1}}(s,a,h) \\
&\geq 1 + \frac{1}{\alpha} \sum_{j: n_{k_j}^{k_{j+1}}(s,a,h) \geq N'_{k_j}(s,a,h)/\alpha} N'_{k_j}(s,a,h) \\
&\geq 1 + \frac{1}{\alpha} \sum_{j=1}^{J(s,a,h)} \left(1+\frac{1}{\alpha}\right)^{j}
\end{aligned}$$

The first inequality follows from focusing only on the epochs where we update due to (s,a,h), where the $+1$ accounts for the first update due to the observing the given state-action-step triple. The second inequality follows from the condition in the subscript of the summation, e.g. we are updating due to (s,a,h). The final inequality follows from the definition of how we trigger updates and because we update $J(s,a,h)$ times due to (s,a,h). Since $\alpha \in [1,\infty)$, Lemma 7 applies. Rearranging terms reveals that:

$$\sum_{j=1}^{J(s,a,h)} \left(1+\frac{1}{\alpha}\right)^{j} \geq \left(1+\frac{1}{\alpha}\right)^{J(s,a,h)+1} - \left(1+\frac{1}{\alpha}\right)$$

Therefore, for $N'_{K+1}(s,a,h) > 0$:

$$\begin{aligned}
N'_{K+1}(s,a,h) &\geq 1 - \frac{1}{\alpha}\left(1+\frac{1}{\alpha}\right) + \frac{1}{\alpha}\left(1+\frac{1}{\alpha}\right)^{J(s,a,h)+1} \\
&> \frac{1}{\alpha}\left(1+\frac{1}{\alpha}\right)^{J(s,a,h)+1} - \frac{1}{\alpha}\left(1+\frac{1}{\alpha}\right)
\end{aligned}$$

If $N'_{K+1}(s,a,h) = 0$ it follows we never update due to this state-action-step triple, which means that $J(s,a,h) = 0$ too. Plugging this into the above expression reveals that:

$$N'_{K+1}(s,a,h) = \frac{1}{\alpha}\left(1+\frac{1}{\alpha}\right)^{J(s,a,h)+1} - \frac{1}{\alpha}\left(1+\frac{1}{\alpha}\right) = 0$$

Thus, for all possible values of the observed visitation counter, we have that:

$$N'_{K+1}(s,a,h) \geq \frac{1}{\alpha}\left(1+\frac{1}{\alpha}\right)^{J(s,a,h)+1} - \frac{1}{\alpha}\left(1+\frac{1}{\alpha}\right)$$

Using the above inequality, we have that

$$\begin{aligned}
T &= \sum_{s,a,h} N_{K+1}(s,a,h) \\
&\geq \sum_{s,a,h} N'_{K+1}(s,a,h) \\
&\geq \sum_{s,a,h} \left(\frac{1}{\alpha}\left(1+\frac{1}{\alpha}\right)^{J(s,a,h)+1} - \frac{1}{\alpha}\left(1+\frac{1}{\alpha}\right)\right) \\
&= -\frac{HSA}{\alpha}\left(1+\frac{1}{\alpha}\right) + \sum_{s,a,h} \frac{1}{\alpha}\left(1+\frac{1}{\alpha}\right)^{J(s,a,h)+1}
\end{aligned}$$

$$\geq -\frac{HSA}{\alpha}\left(1+\frac{1}{\alpha}\right)+\frac{HSA}{\alpha}\left(1+\frac{1}{\alpha}\right)^{\frac{HSA+\sum_{s,a,h}J(s,a,h)}{HSA}} \qquad \text{(Jensen's inequality)}$$
$$\geq -\frac{HSA}{\alpha}\left(1+\frac{1}{\alpha}\right)+\frac{HSA}{\alpha}\left(1+\frac{1}{\alpha}\right)^{\frac{J}{HSA}}$$

where the final line follows from the fact that $J \leq HSA + \sum_{s,a,h} J(s,a,h)$ because we may or may not visit every state-action-step. Rearranging this gives:

$$\frac{T\alpha}{HSA}+1 \geq \left(1+\frac{1}{\alpha}\right)^{\frac{J}{HSA}}$$

Taking logs of both sides and rearranging one last time gives:

$$\begin{aligned} J &\leq HSA \log_{1+1/\alpha}\left(\frac{T\alpha}{HSA}+1\right) \\ &= \frac{HSA\log\left(\frac{T\alpha}{HSA}+1\right)}{\log(1+\frac{1}{\alpha})} \\ &= \frac{HSA\log\left(\frac{K\alpha+SA}{SA}\right)}{\log(1+\frac{1}{\alpha})} \end{aligned}$$

as required. □

Lemma 5. *If $n_0, n_1, \cdots, n_J$ are an arbitrary sequence of real-valued numbers satisfying $n_0 \coloneqq 0$ and $0 \leq n_j \leq \frac{1}{\alpha}N_{j-1}$ with $N_{j-1} = \max\{1, \sum_{i=0}^{j-1} n_i\}$ for all $j \leq J$, then*

$$\sum_{j=1}^{J}\frac{n_j}{N_{j-1}^p} \leq \begin{cases}\left(\sqrt{2}(1+\frac{1}{\alpha})+1\right)\sqrt{N_J} & \text{if } p=\frac{1}{2}\\ (1+\frac{1}{\alpha})+(1+\frac{1}{\alpha})\log\left(N_J\right) & \text{if } p=1\end{cases}$$

Proof. We prove the claim via induction in a similar manner to Jaksch et al. (2010). First, consider the case where $p = 1/2$. Suppose

$$\begin{aligned} \sum_{j=1}^{J-1} n_j \leq 1 &\implies N_1 = N_2 = \cdots = N_{J-1} = 1 && (N_{j-1} = \max\{1, \textstyle\sum_{i=0}^{j-1} n_i\}) \\ &\implies n_J \in [0, N_{J-1}] = \left[0, \frac{1}{\alpha}\right] \end{aligned}$$

Then,

$$\sum_{j=1}^{J}\frac{n_j}{\sqrt{N_{j-1}}} = \sum_{j=1}^{J} n_j = n_J + \sum_{j=1}^{J-1} n_j \leq \frac{1}{\alpha}+1 \leq c\sqrt{N_J} \qquad \text{(For } c \geq 1+1/\alpha)$$

because $N_J \geq 1$. The above is our base case and covers us as long as $\sum_{j=1}^{J-1} n_j \leq 1$ e.g., when $J = 1$ due to $n_0 \coloneqq 1$. Now, we assume the above holds for $\sum_{j=1}^{J-1} n_j > 1$:

$$\sum_{j=1}^{J-1}\frac{n_j}{\sqrt{N_{j-1}}} \leq c\sqrt{N_{J-1}}$$

Finally, we prove the claim holds for J:

$$\begin{aligned}
\sum_{j=1}^{J} \frac{n_j}{\sqrt{N_{j-1}}} &= \frac{n_J}{\sqrt{N_{J-1}}} + \sum_{j=1}^{J-1} \frac{n_j}{\sqrt{N_{j-1}}} \\
&\leq \frac{n_J}{\sqrt{N_{J-1}}} + c\sqrt{N_{J-1}} && \text{(Induction Hypothesis)} \\
&= \sqrt{\left(\frac{n_J}{\sqrt{N_{J-1}}} + c\sqrt{N_{J-1}}\right)^2} \\
&= \sqrt{\frac{n_J^2}{N_{J-1}} + 2cn_J + c^2 N_{J-1}} \\
&\leq \sqrt{\frac{1}{\alpha} n_J + 2cn_J + c^2 N_{J-1}} && \text{(As } n_J \in [0, N_{J-1}/\alpha) \\
&\leq \sqrt{n_J + 2cn_J + c^2 N_{J-1}} && \text{(As } \alpha \geq 1) \\
&= \sqrt{(1+2c)\, n_J + c^2 N_{J-1}} \\
&\leq c\sqrt{n_J + N_{J-1}} && \text{(Pick } c : c^2 \geq 1 + 2c) \\
&= c\sqrt{N_J}
\end{aligned}$$

where the final inequality follows from the fact that $\sum_{j=1}^{J-1} n_j > 1 \implies N_J = n_J + N_{J-1}$. All that remains is selecting c. Using the quadratic formula to find the roots of $c^2 - 2c - 1 = 0$, one can deduce that selecting:

$$c = 1 + \sqrt{2}\left(1 + \frac{1}{\alpha}\right)$$

satisfies $c \geq 1 + 1/\alpha$ and

$$\begin{aligned}
c^2 &= 1 + 2\sqrt{2}\left(1 + \frac{1}{\alpha}\right) + 2\left(1 + \frac{1}{\alpha}\right)^2 \\
&\geq 1 + 2\sqrt{2}\left(1 + \frac{1}{\alpha}\right) + && (\alpha \geq 1) \\
&= 1 + 2\left(1 + \sqrt{2}\left(1 + \frac{1}{\alpha}\right)\right) \\
&= 1 + 2c
\end{aligned}$$

giving the required result. All that remains is to prove the claim for $p = 1$. Similarly to before, suppose:

$$\begin{aligned}
\sum_{j=1}^{J-1} n_j \leq 1 &\implies N_1 = N_2 = \cdots = N_{J-1} = 1 && (N_{j-1} = \max\{1, \textstyle\sum_{i=0}^{j-1} n_i\}) \\
&\implies n_J \in [0, N_{J-1}] = \left[0, \frac{1}{\alpha}\right]
\end{aligned}$$

Then,

$$\sum_{j=1}^{J} \frac{n_j}{N_{j-1}} = \sum_{j=1}^{J} n_j = n_J + \sum_{j=1}^{J-1} n_j \leq \frac{1}{\alpha} + 1 \leq \left(1 + \frac{1}{\alpha}\right) + \left(1 + \frac{1}{\alpha}\right)\log(N_J)$$

because $N_J \geq 1$. The above is our base case and covers us as long as $\sum_{j=1}^{J-1} n_j \leq 1$ e.g., when $J = 1$ due to $n_0 \coloneqq 1$.

Now, we assume the above holds for $\sum_{j=1}^{J-1} n_j > 1$:

$$\sum_{j=1}^{J-1} \frac{n_j}{N_{j-1}} \leq \left(1+\frac{1}{\alpha}\right) + \left(1+\frac{1}{\alpha}\right)\log(N_{J-1})$$

Finally, we prove the claim holds for J:

$$\begin{aligned}
\sum_{j=1}^{J} \frac{n_j}{N_{j-1}} &= \frac{n_J}{N_{J-1}} + \sum_{j=1}^{J-1} \frac{n_j}{N_{j-1}} \\
&\leq \frac{n_J}{N_{J-1}} + \left(1+\frac{1}{\alpha}\right) + \left(1+\frac{1}{\alpha}\right)\log(N_{J-1}) && \text{(Induction Hypothesis)} \\
&\leq \left(1+\frac{1}{\alpha}\right)\log\left(\frac{n_J}{N_{J-1}}+1\right) + \left(1+\frac{1}{\alpha}\right) + \left(1+\frac{1}{\alpha}\right)\log(N_{J-1}) && (n_j/N_{j-1} \in [0,1] \text{ for all } j \leq J) \\
&= \left(1+\frac{1}{\alpha}\right) + \left(1+\frac{1}{\alpha}\right)\log\left(N_{J-1}\left(\frac{n_J}{N_{J-1}}+1\right)\right) \\
&= \left(1+\frac{1}{\alpha}\right) + \left(1+\frac{1}{\alpha}\right)\log\left(n_J + N_{J-1}\right) \\
&= \left(1+\frac{1}{\alpha}\right) + \left(1+\frac{1}{\alpha}\right)\log\left(N_J\right)
\end{aligned}$$

where the final inequality follows from the fact that $\sum_{j=1}^{J-1} n_j > 1 \implies N_{J-1} = \sum_{j=1}^{J-1} n_j$. □

Lemma 7. *Let $\alpha \in [1, \infty)$. Then*

$$\sum_{i=0}^{n} \left(1+\frac{1}{\alpha}\right)^i \geq \left(1+\frac{1}{\alpha}\right)^{n+1} - \frac{1}{\alpha}$$

Proof. Trivially, the statement is true for $n = 0$, because $(1+1/\alpha)^0 = 1$ and $(1+1/\alpha)^1 - 1/\alpha = 1$. Thus, we proceed by induction. Suppose

$$\sum_{i=0}^{n} \left(1+\frac{1}{\alpha}\right)^i \geq \left(1+\frac{1}{\alpha}\right)^{n+1} - \frac{1}{\alpha}$$

for some n. Then

$$\begin{aligned}
\sum_{i=0}^{n+1} \left(1+\frac{1}{\alpha}\right)^i &= \left(1+\frac{1}{\alpha}\right)^{n+1} + \sum_{i=0}^{n} \left(1+\frac{1}{\alpha}\right)^i \\
&\geq \left(1+\frac{1}{\alpha}\right)^{n+1} + \left(1+\frac{1}{\alpha}\right)^{n+1} - \frac{1}{\alpha} \\
&= 2\left(1+\frac{1}{\alpha}\right)^{n+1} - \frac{1}{\alpha} \\
&\geq \left(1+\frac{1}{\alpha}\right)\left(1+\frac{1}{\alpha}\right)^{n+1} - \frac{1}{\alpha} && \text{(Since } 2 \geq 1+1/\alpha\text{)} \\
&= \left(1+\frac{1}{\alpha}\right)^{n+2} - \frac{1}{\alpha}
\end{aligned}$$

Thus, the claim holds for $n+1$, which proves the lemma for all $n \geq 0$. □

Lemma 8. *Algorithm 2 ensures that the summation of the counters across the episodes where we do not update have the following upper bounds:*

$$\sum_{s,a,h} \sum_{j=1}^{J} \frac{n_{k_j+1,h}^{k_{j+1}}(s,a)}{N'_{k_j h}(s,a)} \leq \left(1+\frac{1}{\alpha}\right) HSA + \left(1+\frac{1}{\alpha}\right) HSA \log\left(\frac{K}{SA}\right) \leq 2\left(1+\frac{1}{\alpha}\right) HSA \left(\frac{K}{SA}\right)$$

where the final inequality holds for $K/SA \geq \exp(1)$.

Proof. To prove the result, we extend the summation to include the state-action-step triples in episode k_j that did not trigger the update rule:

$$\begin{aligned}
\sum_{s,a,h}\sum_{j=1}^{J}\frac{n_{k_{j+1},h}^{k_{j+1}}(s,a)}{N'_{k_jh}(s,a)} &\leq \sum_{s,a,h}\sum_{j=1}^{J}\frac{n_{k_jh}^{k_{j+1}}(s,a)}{N'_{k_jh}(s,a)}\mathbb{1}\left\{n_{k_jh}(s,a)\leq\frac{1}{\alpha}N'_{k_jh}(s,a)\right\}\\
&\leq \sum_{s,a,h}\left(\left(1+\frac{1}{\alpha}\right)+\left(1+\frac{1}{\alpha}\right)\log\left(N_J(s,a,h)\right)\right) && \text{(Lemma 5)}\\
&= \left(1+\frac{1}{\alpha}\right)HSA+\left(1+\frac{1}{\alpha}\right)\sum_{s,a,h}\log\left(N_J(s,a,h)\right) && \text{(Expand Summation)}\\
&\leq \left(1+\frac{1}{\alpha}\right)HSA+\left(1+\frac{1}{\alpha}\right)HSA\log\left(\frac{\sum_{s,a,h}N_J(s,a,h)}{HSA}\right) && \text{(Jensen's)}\\
&\leq \left(1+\frac{1}{\alpha}\right)HSA+\left(1+\frac{1}{\alpha}\right)HSA\log\left(\frac{T}{HSA}\right) && \left(\textstyle\sum_{s,a,h}N_J(s,a,h)\leq T\right)\\
&= \left(1+\frac{1}{\alpha}\right)HSA+\left(1+\frac{1}{\alpha}\right)HSA\log\left(\frac{K}{SA}\right) && (T=KH)\\
&\leq 2\left(1+\frac{1}{\alpha}\right)HSA\log\left(\frac{K}{SA}\right)
\end{aligned}$$

for $K/SA\geq\exp(1)$, as required. □

A.4 Proof of Regret Bound for Lazy Updating

Theorem 2. *Let $K\geq SA$ and $\alpha\geq 1$. Under Assumption 1 and 2, with probability $1-\delta$, the regret of any model-based algorithm under delayed feedback is upper bounded by:*

$$\mathfrak{R}_K\lesssim\left(1+\frac{1}{\alpha}\right)\hat{\mathfrak{R}}_K(Base)+\frac{H^2SA\mathbb{E}[\tau]}{\log(1+\frac{1}{\alpha})}$$

where $\hat{\mathfrak{R}}_K(Base)$ is an upper bound on the regret of the chosen base algorithm under immediate feedback.

Proof. Let $\tilde{\Delta}_h^k(s)=\tilde{V}_h^{\pi_k}(s)-V_h^{\pi_k}(s)$ denote the difference between the optimistic and actual value of policy π_k from state s and step h. By definition, the regret of any episodic reinforcement learning algorithm is given by:

$$\begin{aligned}
\mathfrak{R}_K &= \sum_{k=1}^{K}\left(V_1^*\left(s_1^k\right)-V_1^{\pi_k}\left(s_1^k\right)\right)\\
&\leq \sum_{k=1}^{K}\left(\tilde{V}_1^{\pi_k}\left(s_1^k\right)-V_1^{\pi_k}\left(s_1^k\right)\right)=\sum_{k=1}^{K}\tilde{\Delta}_1^k\left(s_1^k\right)=\sum_{j=1}^{J}\sum_{k=k_j}^{k_{j+1}-1}\tilde{\Delta}_1^k\left(s_1^k\right)\\
&= \sum_{j=1}^{J}\sum_{k=k_j}^{k_{j+1}-1}\tilde{\Delta}_1^k\left(s_1^k\right)\mathbb{1}\left\{k+\tau_k\geq k_{j+1}\right\}=\sum_{j=1}^{J}\tilde{\Delta}_1^{k_j}\left(s_1^{k_j}\right)+\sum_{j=1}^{J}\sum_{k=k_j}^{k_{j+1}-1}\tilde{\Delta}_1^k\left(s_1^k\right)\\
&= \underbrace{\sum_{j=1}^{J}\tilde{\Delta}_1^{k_j}\left(s_1^{k_j}\right)}_{(i)}+\underbrace{\sum_{j=1}^{J}\sum_{k=k_j+1}^{k_{j+1}-1}\tilde{\Delta}_1^k\left(s_1^k\right)\mathbb{1}\left\{k+\tau_k\geq k_{j+1}\right\}}_{(ii)}+\underbrace{\sum_{j=1}^{J}\sum_{k=k_j+1}^{k_{j+1}-1}\tilde{\Delta}_1^k\left(s_1^k\right)\mathbb{1}\left\{k+\tau_k<k_{j+1}\right\}}_{(iii)}
\end{aligned}$$

where the inequality follows from optimism, the penultimate equality follows from epochs consisting of disjoint sets of episodes and the final equality follows from splitting the episodes into three disjoint sets, (i), (ii), and (ii):

(i) episodes where we perform a policy update,

(ii) episodes played in the j-th epoch but observed in epoch $j'>j$,

(iii) episodes played in the j-th epoch and observed in the j-th epoch.

First, we focus on the episodes where we perform a policy update, e.g. (i). Recall that Lemma 4 tells us the total number of updates is logarithmic in the number of episodes. Further, the rewards are bounded between zero and one, meaning the regret of any episode is at most H. Combining these two results gives a trivial bound on regret of this term: $(i) \leq HJ$.

Next, we bound the regret of the episodes whose feedback is not observable before the start of the next epoch e.g., (ii). Once again, we can rely on Lemma 4 and the fact that the regret of any episode is at most H to get a bound on this term that is logarithmic in K. Doing so gives the following result:

$$\begin{aligned}
(ii) &= \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \tilde{\Delta}_1^k\left(s_1^k\right) \mathbb{1}\left\{k+\tau_k \geq k_{j+1}\right\} \\
&\leq H \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \mathbb{1}\left\{k+\tau_k \geq k_{j+1}\right\} \\
&\leq H \sum_{j=1}^{J} \sum_{k=1}^{k_{j+1}-1} \mathbb{1}\left\{k+\tau_k \geq k_{j+1}\right\} \\
&= H \sum_{j=1}^{J} S_{k_{j+1}} \\
&\leq HJ\psi_K^\tau && (S_k \leq \psi_k^\tau \leq \psi_K^\tau)
\end{aligned}$$

Finally, we handle the episodes that are played and observed in the same epoch e.g., term (iii). Lemma 2 allows us to make a start on bounding this term:

$$\begin{aligned}
(iii) &= \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \tilde{\Delta}_1^k\left(s_1^k\right) \mathbb{1}\left\{k+\tau_k < k_{j+1}\right\} \\
&\leq 6\left(H+C\right)\sqrt{T\log\left(\frac{K\pi}{6\delta'}\right)} + 6\sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} \left(\beta_{kh}\left(s_h^k, a_h^k\right) + \frac{3CH^2SL}{N'_{kh}\left(s_h^k, a_h^k\right)}\right) \mathbb{1}\left\{k+\tau_k < k_{j+1}\right\} \\
&= 6\left(H+C\right)\sqrt{T\log\left(\frac{K\pi}{6\delta'}\right)} \\
&\quad + 6\sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} \frac{3CH^2SL}{N'_{kh}\left(s_h^k, a_h^k\right)} \mathbb{1}\left\{k+\tau_k < k_{j+1}\right\} && (iii.a) \\
&\quad + 6\sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} \beta_{kh}\left(s_h^k, a_h^k\right) \mathbb{1}\left\{k+\tau_k < k_{j+1}\right\} && (iii.b)
\end{aligned}$$

Thus, bounding (iii) now amounts to finding an upper bounds for $(iii.a)$ and $(iii.b)$. Since k_j does not feature in either summation, we know that

$$n_{k'h}^{k_{j+1}}(s,a) \leq \frac{1}{\alpha} N'_{k_j h}(s,a)$$

for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $k' \geq k_j + 1$. By introducing a summation over all the states-actions and steps, we can easily bound $(iii.a)$ via Lemma 8:

$$\begin{aligned}
(iii.a) &= 3CH^2SL \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} \frac{\mathbb{1}\left\{k+\tau_k < k_{j+1}\right\}}{N'_{k_j h}\left(s_h^k, a_h^k\right)} \\
&= 3CH^2SL \sum_{s,a,h} \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \frac{\mathbb{1}\left\{s_h^k = s, a_h^k = a, k+\tau_k < k_{j+1}\right\}}{N'_{k_j h}\left(s,a\right)}
\end{aligned}$$

$$
\begin{aligned}
&= \left(B_2 + 3CH^2SL\right) \sum_{s,a,h} \sum_{j=1}^{J} \frac{\sum_{k=k_j+1}^{k_{j+1}-1} \mathbb{1}\left\{s_h^k = s, a_h^k = a, k + \tau_k < k_{j+1}\right\}}{N'_{k_j h}(s,a)} \\
&= 3CH^2SL \sum_{s,a,h} \sum_{j=1}^{J} \frac{n_{k_j+1,h}^{k_{j+1}}(s,a)}{N'_{k_j h}(s,a)} && \text{(Definition in Equation (11))} \\
&\leq 3CH^2SL \left(2\left(1+\frac{1}{\alpha}\right) HSA \log\left(\frac{K}{SA}\right)\right) && \text{(By Lemma 8)} \\
&= 6\left(1+\frac{1}{\alpha}\right) CH^3S^2AL \log\left(\frac{K}{SA}\right)
\end{aligned}
$$

Bounding $(iii.b)$ requires some care due to the various forms of B_1 e.g., those that remain constant and those that utilise variance reduction techniques. By Lemma 5, it is clear that the summation of the visitation counters no longer depends on the delay. Therefore, we begin by an application of Cauchy-Schwarz (CS) to separate the numerator of the exploration bonus from the summation of the visitation counters:

$$
\begin{aligned}
(iii.b) &= \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} \left(\frac{B_1}{\sqrt{N'_{k_j h}\left(s_h^k, a_h^k\right)}} + \frac{B_2}{N'_{k_j h}\left(s_h^k, a_h^k\right)}\right) \mathbb{1}\left\{k + \tau_k < k_{j+1}\right\} \\
&= \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} \frac{B_1 \mathbb{1}\left\{k + \tau_k < k_{j+1}\right\}}{\sqrt{N'_{k_j h}\left(s_h^k, a_h^k\right)}} + B_2 \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} \frac{\mathbb{1}\left\{k + \tau_k < k_{j+1}\right\}}{N'_{k_j h}\left(s_h^k, a_h^k\right)} \\
&\leq \sqrt{\sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} B_1^2 \mathbb{1}\left\{k + \tau_k < k_{j+1}\right\} \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} \frac{\mathbb{1}\left\{k + \tau_k < k_{j+1}\right\}}{N'_{k_j h}\left(s_h^k, a_h^k\right)}} \\
&\quad + B_2 \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} \frac{\mathbb{1}\left\{k + \tau_k < k_{j+1}\right\}}{N'_{k_j h}\left(s_h^k, a_h^k\right)} \\
&= \sqrt{\sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} B_1^2 \mathbb{1}\left\{k + \tau_k < k_{j+1}\right\} \sum_{s,a,h} \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \frac{\mathbb{1}\left\{s_h^k = s, a_h^k = a, k + \tau_k < k_{j+1}\right\}}{N'_{k_j h}\left(s_h^k, a_h^k\right)}} \\
&\quad + B_2 \sum_{s,a,h} \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \frac{\mathbb{1}\left\{s_h^k = s, a_h^k = a, k + \tau_k < k_{j+1}\right\}}{N'_{k_j h}\left(s_h^k, a_h^k\right)} \\
&= \sqrt{\sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} B_1^2 \mathbb{1}\left\{k + \tau_k < k_{j+1}\right\} \sum_{s,a,h} \sum_{j=1}^{J} \frac{\sum_{k=k_j+1}^{k_{j+1}-1} \mathbb{1}\left\{s_h^k = s, a_h^k = a, k + \tau_k < k_{j+1}\right\}}{N'_{k_j h}\left(s_h^k, a_h^k\right)}} \\
&\quad + B_2 \sum_{s,a,h} \sum_{j=1}^{J} \frac{\sum_{k=k_j+1}^{k_{j+1}-1} \mathbb{1}\left\{s_h^k = s, a_h^k = a, k + \tau_k < k_{j+1}\right\}}{N'_{k_j h}\left(s_h^k, a_h^k\right)} \\
&= \sqrt{\sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} B_1^2 \mathbb{1}\left\{k + \tau_k < k_{j+1}\right\} \sum_{s,a,h} \sum_{j=1}^{J} \frac{n_{k_j+1,h}^{k_{j+1}}(s,a)}{N'_{k_j h}\left(s_h^k, a_h^k\right)}} + B_2 \sum_{s,a,h} \sum_{j=1}^{J} \frac{n_{k_j+1,h}^{k_{j+1}}(s,a)}{N'_{k_j h}\left(s_h^k, a_h^k\right)} && \text{(Eq. (11))} \\
&\leq \sqrt{2\left(1+\frac{1}{\alpha}\right) HSA \log\left(\frac{K}{SA}\right) \sum_{j=1}^{J} \sum_{k=k_j+1}^{k_{j+1}-1} \sum_{h=1}^{H} B_1^2 \mathbb{1}\left\{k + \tau_k < k_{j+1}\right\}} \\
&\quad + 2\left(1+\frac{1}{\alpha}\right) B_2 HSA \log\left(\frac{K}{SA}\right) && \text{(Lemma 8)}
\end{aligned}
$$

$$\leq \left(1+\frac{1}{\alpha}\right)\hat{\mathfrak{R}}_K(Base)\log\left(\frac{K}{SA}\right)$$
$$\lesssim \left(1+\frac{1}{\alpha}\right)\hat{\mathfrak{R}}_K(Base)$$

The penultimate line in the above is simply the sum of the bonuses for the chosen base algorithm under immediate feedback scaled by a logarithmic factor, which is introduced by the slower updating. For $B_1 \approx B$ e.g., the upper bound only involves inflating terms inside logarithms, one can upper bound the summation under the square-root by TB^2, which is tight up to logarithmic factors. When B_1 involves some form of empirical variance term, one can use the techniques outlined by Neu and Pike-Burke (2020); Azar et al. (2017); Fruit et al. (2020) to bound the summation under the square-root by $\approx HT$; once again this too is tight up to logarithmic factors. More simply, the epochs form a simulated non-delayed version of the environment for the base algorithm. Therefore, $(iii.b)$ can be replaced with the upper bound of the regret in the non-delayed environment multiplied by the extra logarithmic factors that arise from the slower updating, because the summation of the bonuses are the leading term in the regret bound.

Bringing everything together gives:

$$\begin{aligned}\mathfrak{R}_K &\leq (i)+(ii)+(iii.a)+(iii.b)\\ &\lesssim \left(1+\frac{1}{\alpha}\right)\hat{\mathfrak{R}}_K(Base)+HJ\psi_K^\tau\\ &\lesssim \left(1+\frac{1}{\alpha}\right)\hat{\mathfrak{R}}_K(Base)+\frac{H^2SA\psi_K^\tau}{\log(1+\frac{1}{\alpha})}\end{aligned}$$

Plugging in ψ_K^τ (and suppressing poly-logarithmic factors) gives the stated result. ■

B ADDITIONAL THEORETICAL RESULTS

Here, we present a brief overview of the results that unify model-optimistic and value-optimistic model-based episodic reinforcement learning algorithms (Neu and Pike-Burke, 2020). The class of model-optimistic algorithms explicitly define the following failure event for some divergence $D(\hat{P}_k h(\cdot|s,a), P_h(\cdot|s,a)$:

$$F_k^p = \left\{\exists\, s,a,h : D\left(\hat{P}_{kh}\left(\cdot|s,a\right), P_h\left(\cdot|s,a\right)\right) \geq \epsilon_{kh}^p\left(s,a\right)\right\}$$

which holds across all episodes with probability δ'. Indeed, D must satisfy some conditions. Namely, D must be jointly convex in its arguments so that $\mathcal{P}_{kh}$ (defined below) is convex, and it must be positive homogeneous.[6] Outside the failure event, with probability $1-\delta'$, the divergence between the empirical and actual transition density of the h^{th} step at the start of the k^{th} episode is therefore, at most: $D(\hat{P}_{kh}(\cdot|s,a), P_h(\cdot|s,a)) \leq \epsilon_{kh}^p(s,a)$. Using $\epsilon_{kh}^p(s,a)$ as the maximum divergence allows for the construction of the following plausible set:

$$\mathcal{P}_{kh} = \left\{\tilde{P}_h\left(\cdot|s,a\right) \in \Delta : D\left(\tilde{P}_h\left(\cdot|s,a\right), \hat{P}_{kh}\left(\cdot|s,a\right)\right) \leq \epsilon_{kh}^p\left(s,a\right)\right\}$$

for each $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$. Here, Δ denotes the set of valid transition densities. From here, it is possible to derive the bonus by finding the conjugate of the divergence:

$$\begin{aligned}\beta_{kh}^*\left(s,a\right) &= \max_{\tilde{P}_h(\cdot|s,a)\in\Delta}\left\{\left\langle \tilde{V}, \tilde{P}_h\left(\cdot|s,a\right) - \hat{P}_h\left(\cdot|s,a\right)\right\rangle\right\}\\ \beta_{kh}^-\left(s,a\right) &= \max_{\tilde{P}_h(\cdot|s,a)\in\Delta}\left\{\left\langle -\tilde{V}, \tilde{P}_h\left(\cdot|s,a\right) - \hat{P}_h\left(\cdot|s,a\right)\right\rangle\right\}\\ \beta_{kh}\left(s,a\right) &\geq \max\left\{\beta_{kh}^*\left(s,a\right), \beta_{kh}^-\left(s,a\right)\right\}\end{aligned}$$

by introducing a Lagrange multiplier. For a derivation of the bonuses associated with each divergence, we refer the reader to Appendix A.5 of Neu and Pike-Burke (2020).

[6] The distance $\|p-p'\|$ for any norm and all f -divergences satisfy these conditions (Liese and Vajda, 2006).

B.1 Missing Proofs for the Regret Decomposition

In this subsection, we utilise the fact that all model-based algorithms compute an optimistic value function of the form (4) to derive an adaptable regret decomposition. The decomposition is adaptable in the sense it allows for tighter delay-dependence when the bonuses satisfy a symmetry-like property.

Throughout, we assume that the model-based algorithm is optimistic with high probability. That is, $\tilde{V}_h^{\pi_k}(s) \geq V_h^*(s) \geq V_h^{\pi_k}(s)$ with high probability at least $1-\delta'$. Further, C is defined as the event where:

$$\beta_{kh}^+(s,a)) \geq \left\langle \left(\hat{P}_{kh} - P_h\right)(\cdot\,|s,a)\,, \tilde{V}_{h+1}^{\pi_k}(\cdot)\right\rangle$$

which holds across all episodes for every state-action-step triple conditional on the complement of the failure event.

Lemma 2. *Under Assumption 1, with probability $1-4\delta'$, we can upper bound the regret by:*

$$\begin{aligned}\mathfrak{R}_K &\leq 6\,(H+C)\sqrt{T\log\left(\frac{K\pi}{6\delta'}\right)} \\ &\quad + 6\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_{kh}^+\left(s_h^k,a_h^k\right) + 6\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{3CH^2SL}{N'_{kh}\left(s_h^k,a_h^k\right)}\end{aligned}$$

where $L = \log\left(S^2AH\pi^2/6\delta'\right)$ and C indicates whether the bonuses of the algorithm satisfy Equation (6)*.*

Proof. By definition, the regret of any episodic reinforcement learning algorithm is given by:

$$\begin{aligned}\mathfrak{R}_K &= \sum_{k=1}^{K} V_1^*\left(s_1^k\right) - V_1^{\pi_k}\left(s_1^k\right) \\ &\leq \sum_{k=1}^{K} \tilde{V}_1^{\pi_k}\left(s_1^k\right) - V_1^{\pi_k}\left(s_1^k\right)\end{aligned}$$

where the final inequality holds by optimism, which holds across all episodes with probability at least $1-\delta'$. Consider the more general case of bounding the regret from the h-th step of each episode, rather than just the first step. Define $\tilde{\Delta}_h^k = \tilde{V}_h^{\pi_k}(s_h^k) - V_h^{\pi_k}(s_h^k)$. Applying Lemma 9 gives, with probability at least $1-\delta'$:

$$\tilde{\Delta}_h^k(s_h^k) \leq \left(1+\frac{C}{H}\right)\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) + 2\beta_{kh}\left(s_h^k,a_h^k\right) + \frac{6CH^2SL}{N'_{kh}\left(s_h^k,a_h^k\right)} + \zeta_{h+1}^k + C\bar{\zeta}_{h+1}^k$$

where

$$\begin{aligned}\zeta_{h+1}^k &\coloneqq \left\langle P_h\left(\cdot\,|\,s_h^k,a_h^k\right), \tilde{\Delta}_{h+1}^k(\cdot)\right\rangle - \tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) \\ \bar{\zeta}_{h+1}^k &\coloneqq \sqrt{\frac{4L}{N'_{kh}\left(s_h^k,a_h^k\right)}}\left[\left(\sum_{s'\in G_{kh}} P_h\left(s'\,|\,s_h^k,a_h^k\right)\frac{\tilde{\Delta}_{h+1}^k(s')}{\sqrt{P_h\left(s'\,|\,s_h^k,a_h^k\right)}}\right) - \frac{\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)}{\sqrt{P_h\left(s_{h+1}^k\,|\,s_h^k,a_h^k\right)}}\right]\end{aligned}$$

and

$$G_{kh} \coloneqq \left\{s' : P_h\left(s'\,|\,s_h^k,a_h^k\right)N'_{kh}\left(s_h^k,a_h^k\right) \geq 4H^2L\right\}$$

Now, we can utilise the recursive decomposition above to show that:

$$\tilde{\Delta}_j^k\left(s_j^k\right) \leq \left(1+\frac{C}{H}\right)^{H-j}\sum_{h=j}^{H} 2\beta_{kh}\left(s_h^k,a_h^k\right) + \frac{6CH^2SL}{N'_{kh}\left(s_h^k,a_h^k\right)} + \zeta_{h+1}^k + C\bar{\zeta}_{h+1}^k$$

which we do by induction. Recall that: $\tilde{V}^{\pi_k}_{H+1} = V^*_{H+1} = V^{\pi_k}_{H+1} = \vec{0}$. Therefore, the statement holds when $j = H$, because: $\tilde{\Delta}^k_{H+1} = 0$. Now assume the statement holds for $h = j+1$. Then,

$$\begin{aligned}
\tilde{\Delta}^k_j\left(s^k_j\right) &\leq \left(1+\frac{C}{H}\right)\tilde{\Delta}^k_{j+1}\left(s^k_{j+1}\right) + \left(2\beta_{kh}\left(s^k_h,a^k_h\right) + \frac{6CH^2SL}{N'_{kh}\left(s^k_h,a^k_h\right)} + \zeta^k_{h+1} + C\bar{\zeta}^k_{h+1}\right) \\
&\leq \left(1+\frac{C}{H}\right)\left(\left(1+\frac{C}{H}\right)^{H-(j+1)}\sum_{h=j+1}^{H} 2\beta_{kh}\left(s^k_h,a^k_h\right) + \frac{6CH^2SL}{N'_{kh}\left(s^k_h,a^k_h\right)} + \zeta^k_{h+1} + C\bar{\zeta}^k_{h+1}\right) \\
&\quad + \left(2\beta_{kh}\left(s^k_h,a^k_h\right) + \frac{6CH^2SL}{N'_{kh}\left(s^k_h,a^k_h\right)} + \zeta^k_{h+1} + C\bar{\zeta}^k_{h+1}\right) \\
&= \left(1+\frac{C}{H}\right)^{H-j}\sum_{h=j+1}^{H}\left(2\beta_{kh}\left(s^k_h,a^k_h\right) + \frac{6CH^2SL}{N'_{kh}\left(s^k_h,a^k_h\right)} + \zeta^k_{h+1} + C\bar{\zeta}^k_{h+1}\right) \\
&\quad + \left(2\beta_{kh}\left(s^k_h,a^k_h\right) + \frac{6CH^2SL}{N'_{kh}\left(s^k_h,a^k_h\right)} + \zeta^k_{h+1} + C\bar{\zeta}^k_{h+1}\right) \\
&\leq \left(1+\frac{C}{H}\right)^{H-j}\sum_{h=j+1}^{H}\left(2\beta_{kh}\left(s^k_h,a^k_h\right) + \frac{6CH^2SL}{N'_{kh}\left(s^k_h,a^k_h\right)} + \zeta^k_{h+1} + C\bar{\zeta}^k_{h+1}\right) \\
&\quad + \left(1+\frac{C}{H}\right)^{H-j}\left(2\beta_{kh}\left(s^k_h,a^k_h\right) + \frac{6CH^2SL}{N'_{kh}\left(s^k_h,a^k_h\right)} + \zeta^k_{h+1} + C\bar{\zeta}^k_{h+1}\right) \\
&\leq \left(1+\frac{C}{H}\right)^{H-j}\sum_{h=j}^{H}\left(2\beta_{kh}\left(s^k_h,a^k_h\right) + \frac{6CH^2SL}{N'_{kh}\left(s^k_h,a^k_h\right)} + \zeta^k_{h+1} + C\bar{\zeta}^k_{h+1}\right)
\end{aligned}$$

Therefore, we are now able to upper bound the regret as follows:

$$\begin{aligned}
\mathfrak{R}_K &\leq \sum_{k=1}^{K}\tilde{\Delta}^k_1\left(s^k_1\right) \\
&\leq \underbrace{\left(1+\frac{C}{H}\right)^{H}}_{\leq e<3}\left(C\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\zeta}^k_{h+1} + \sum_{k=1}^{K}\sum_{h=1}^{H}\zeta^k_{h+1} + 2\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_{kh}\left(s^k_h,a^k_h\right) + \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{6CH^2SL}{N'_{kh}\left(s^k_h,a^k_h\right)}\right) \\
&\leq 3C\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\zeta}^k_{h+1} + 3\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta^k_{h+1} + 6\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_{kh}\left(s^k_h,a^k_h\right) + \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{18CH^2SL}{N'_{kh}\left(s^k_h,a^k_h\right)} \\
&= 3C\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\zeta}^k_{h+1} + 3\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta^k_{h+1} + 6\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\beta_{kh}\left(s^k_h,a^k_h\right) + \frac{3CH^2SL}{N'_{kh}\left(s^k_h,a^k_h\right)}\right)
\end{aligned}$$

Recall the definitions of ζ^k_{h+1} and $\bar{\zeta}^k_{h+1}$:

$$\begin{aligned}
\zeta^k_{h+1} &= \left\langle P_h\left(\cdot \mid s^k_h,a^k_h\right), \tilde{\Delta}^k_{h+1}\left(\cdot\right)\right\rangle - \tilde{\Delta}^k_{h+1}(s^k_{h+1}) \\
\bar{\zeta}^k_{h+1} &= \sqrt{\frac{4L}{N'_{kh}\left(s^k_h,a^k_h\right)}}\left[\left(\sum_{s'\in G_{kh}}\frac{P_h\left(s'|s^k_h,a^k_h\right)\tilde{\Delta}^k_{h+1}\left(s'\right)}{\sqrt{P_h\left(s'|s^k_h,a^k_h\right)}}\right) - \frac{\Delta^k_{h+1}\left(s^k_{h+1}\right)}{\sqrt{P_h\left(s^k_{h+1}|s^k_h,a^k_h\right)}}\right]
\end{aligned}$$

with

$$G_{kh} \coloneqq \{s' : P_h\left(s' \mid s^k_h,a^k_h\right)N'_{kh}\left(s^k_h,a^k_h\right) \geq 4H^2L\}$$

Let $\mathcal{F}_{kh} = \sigma(\{\mathcal{H}_i\}_{i:i+\tau_i<k})$ be the natural filtration of the observed information. Then $|\zeta_{h+1}^k| \le 2H$ and

$$\begin{aligned}
&\mathbb{E}_{s_{h+1}^k \sim P_h(\cdot \mid s_h^k, a_h^k)} \left[\zeta_{h+1}^k \mid \mathcal{F}_{kh} \cup \{s_h^k, a_h^k\}\right] \\
&= \mathbb{E}_{s_{h+1}^k \sim P_h(\cdot \mid s_h^k, a_h^k)} \left[\left\langle P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k\left(\cdot\right)\right\rangle - \tilde{\Delta}_{h+1}^k\left(s_{h+1}\right) \,\middle|\, \mathcal{F}_{kh} \cup \{s_h^k, a_h^k\}\right] \\
&= \left\langle P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k\left(\cdot\right)\right\rangle - \mathbb{E}_{s_{h+1}^k \sim P_h(\cdot \mid s_h^k, a_h^k)} \left[\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) \,\middle|\, \mathcal{F}_{kh} \cup \{s_h^k, a_h^k\}\right] \\
&= \left\langle P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k\left(\cdot\right)\right\rangle - \left\langle P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k\left(\cdot\right)\right\rangle = 0
\end{aligned}$$

Similarly, $|\bar{\zeta}_{h+1}^k| \le 2$ and $\mathbb{E}_{s_{h+1}^k \sim P_h(\cdot \mid s_h^k, a_h^k)}\left[\bar{\zeta}_{h+1}^k \mid \mathcal{F}_{kh} \cup \{s_h^k, a_h^k\}, s_{h+1}^k \in G_{kh}\right] = 0$. Therefore, ζ_{h+1}^k and $\bar{\zeta}_{h+1}^k$ are martingale differences, which are easily bounded using Azuma-Hoeffding:

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_{h+1}^k \le 2H\sqrt{T\log\left(\frac{K\pi}{6\delta'}\right)} \qquad \text{(with probability at least } 1-\delta')$$

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \bar{\zeta}_{h+1}^k \le 2\sqrt{T\log\left(\frac{K\pi}{6\delta'}\right)} \qquad \text{(with probability at least } 1-\delta')$$

Therefore, with probability $1-4\delta'$:

$$\begin{aligned}
\mathfrak{R}_K &\le 6C\sqrt{T\log\left(\frac{K\pi}{6\delta'}\right)} + 6H\sqrt{T\log\left(\frac{K\pi}{6\delta'}\right)} + 6\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\beta_{kh}\left(s_h^k, a_h^k\right) + \frac{3CH^2SL}{N'_{kh}\left(s_h^k, a_h^k\right)}\right) \\
&\le 6(H+C)\sqrt{T\log\left(\frac{K\pi}{6\delta'}\right)} + 6\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\beta_{kh}\left(s_h^k, a_h^k\right) + \frac{3CH^2SL}{N'_{kh}\left(s_h^k, a_h^k\right)}\right)
\end{aligned}$$

as required. □

Lemma 9. *Let C be an algorithm dependent-constant indicating whether it is model-optimistic or value-optimistic. Under Assumption 1, the regret of any optimistic model-based algorithm from the h-th step of the k-th episode upper bounded by:*

$$\begin{aligned}
\tilde{\Delta}_h^k(s_h^k) \le{}& \left(1+\frac{C}{H}\right)\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) + 2\beta_{kh}\left(s_h^k, a_h^k\right) + \frac{6CH^2SL}{N'_{kh}\left(s_h^k, a_h^k\right)} \\
&+ \left\langle P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k\left(\cdot\right)\right\rangle - \tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) \\
&+ \sqrt{\frac{4CL}{N'_{kh}\left(s_h^k, a_h^k\right)}}\left[\left(\sum_{s'\in G_{kh}} P_h\left(s' \mid s_h^k, a_h^k\right)\frac{\tilde{\Delta}_{h+1}^k\left(s'\right)}{\sqrt{P_h\left(s' \mid s_h^k, a_h^k\right)}}\right) - \frac{\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)}{\sqrt{P_h\left(s_{h+1}^k \mid s_h^k, a_h^k\right)}}\right]
\end{aligned}$$

where $L = \log(S^2AH\pi^2/6\delta')$ and

$$G_{kh} \coloneqq \left\{s' : P_h\left(s' \mid s_h^k, a_h^k\right) N'_{kh}\left(s_h^k, a_h^k\right) \ge 4H^2L\right\}$$

with probability $1-\delta'$.

Proof. By Proposition 2 of Neu and Pike-Burke (2020) and by definition of the value-optimistic algorithms, we have that:

$$\begin{aligned}
&\tilde{\Delta}_h^k(s_h^k) = \tilde{V}_h^{\pi_k}\left(s_h^k\right) - V_h^{\pi_k}\left(s_h^k\right) \\
&= \beta_{kh}^+\left(s_h^k, a_h^k\right) + \left\langle \hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right)\tilde{V}_{h+1}^{\pi_k}\right\rangle - \left\langle P_h\left(\cdot \mid s_h^k, a_h^k\right)V_{h+1}^{\pi_k}\right\rangle \\
&= \beta_{kh}^+\left(s_h^k, a_h^k\right) + \left\langle \hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{V}_{h+1}^{\pi_k}\right\rangle + \left\langle P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{V}_{h+1}^{\pi_k} - V_{h+1}^{\pi_k}\right\rangle \\
&= \beta_{kh}^+\left(s_h^k, a_h^k\right) + \left\langle \hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{V}_{h+1}^{\pi_k}\right\rangle + \left\langle P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k\right\rangle \\
&\le \beta_{kh}\left(s_h^k, a_h^k\right) + \left\langle \hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{V}_{h+1}^{\pi_k}\right\rangle + \left\langle P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k\right\rangle \\
&= \tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) + \beta_{kh}\left(s_h^k, a_h^k\right) + \left\langle \hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{V}_{h+1}^{\pi_k}\right\rangle + \left\langle P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k\right\rangle - \tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)
\end{aligned}$$

where the inequality follows from the fact that $\beta_{kh}(s,a)^+ \leq \beta_{kh}(s,a)$. For model-optimistic algorithms, from the definition of the bonuses, we have that:

$$\left\langle \hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{V}_{h+1}^{\pi_k} \right\rangle \leq \beta_{kh}^-\left(s_h^k, a_h^k\right) \leq \beta_{kh}\left(s_h^k, a_h^k\right)$$

However, this term cannot be bound as easily for the value-optimistic algorithms. But, Assumption 1 allows us to show that:

$$\begin{aligned}
&\left\langle \hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{V}_{h+1}^{\pi_k} \right\rangle \\
&= \left\langle \hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right), V_{h+1}^* \right\rangle + \left\langle \hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{V}_{h+1}^{\pi_k} - V_{h+1}^* \right\rangle \\
&\leq \beta_{kh}\left(s_h^k, a_h^k\right) + \left\langle \hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{V}_{h+1}^{\pi_k} - V_{h+1}^* \right\rangle && \text{(By Assumption 1)} \\
&\leq \beta_{kh}\left(s_h^k, a_h^k\right) + \left\langle |\hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right)|, \tilde{V}_{h+1}^{\pi_k} - V_{h+1}^* \right\rangle \\
&\leq \beta_{kh}\left(s_h^k, a_h^k\right) + \left\langle |\hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right)|, \tilde{V}_{h+1}^{\pi_k} - V_{h+1}^{\pi_k} \right\rangle && (V_h^*(s) \geq V_h^{\pi_k}(s)) \\
&= \beta_{kh}\left(s_h^k, a_h^k\right) + \left\langle |\hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right)|, \tilde{\Delta}_{h+1}^k \right\rangle \\
&\leq \beta_{kh}\left(s_h^k, a_h^k\right) + \frac{\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)}{H} + \frac{2HSL}{N'_{kh}\left(s_h^k, a_h^k\right)} + \frac{4H^2SL}{N'_{kh}\left(s_h^k, a_h^k\right)} \\
&+ \sqrt{\frac{4L}{N'_{kh}\left(s_h^k, a_h^k\right)}} \left[\left(\sum_{s' \in G_{kh}} P_h\left(s' \mid s_h^k, a_h^k\right) \frac{\tilde{\Delta}_{h+1}^k(s')}{\sqrt{P_h\left(s' \mid s_h^k, a_h^k\right)}} \right) - \frac{\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)}{\sqrt{P_h\left(s_{h+1}^k \mid s_h^k, a_h^k\right)}} \right] && \text{(By Lemma 10)}
\end{aligned}$$

with probability $1-\delta'$. Thus, utilising the indicator variable, we have that:

$$\begin{aligned}
\tilde{\Delta}_h^k(s_h^k) &\leq \left(1 + \frac{C}{H}\right) \tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) + 2\beta_{kh}\left(s_h^k, a_h^k\right) + \frac{2CHSL}{N'_{kh}\left(s_h^k, a_h^k\right)} + \frac{4CH^2SL}{N'_{kh}\left(s_h^k, a_h^k\right)} \\
&\quad + \left\langle P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k(\cdot) \right\rangle - \tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) \\
&\quad + \sqrt{\frac{4CL}{N'_{kh}\left(s_h^k, a_h^k\right)}} \left[\left(\sum_{s' \in G_{kh}} P_h\left(s' \mid s_h^k, a_h^k\right) \frac{\tilde{\Delta}_{h+1}^k(s')}{\sqrt{P_h\left(s' \mid s_h^k, a_h^k\right)}} \right) - \frac{\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)}{\sqrt{P_h\left(s_{h+1}^k \mid s_h^k, a_h^k\right)}} \right] \\
&\leq \left(1 + \frac{C}{H}\right) \tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) + 2\beta_{kh}\left(s_h^k, a_h^k\right) + \frac{6CH^2SL}{N'_{kh}\left(s_h^k, a_h^k\right)} \\
&\quad + \left\langle P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k(\cdot) \right\rangle - \tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) \\
&\quad + \sqrt{\frac{4CL}{N'_{kh}\left(s_h^k, a_h^k\right)}} \left[\left(\sum_{s' \in G_{kh}} P_h\left(s' \mid s_h^k, a_h^k\right) \frac{\tilde{\Delta}_{h+1}^k(s')}{\sqrt{P_h\left(s' \mid s_h^k, a_h^k\right)}} \right) - \frac{\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)}{\sqrt{P_h\left(s_{h+1}^k \mid s_h^k, a_h^k\right)}} \right]
\end{aligned}$$

as required. □

Lemma 10. *Let* $\gamma_{kh}(s_h^k, a_h^k) \coloneqq \langle \hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k \rangle$. *Then, with probability at least* $1-\delta'$*:*

$$\begin{aligned}
\gamma_{kh}\left(s_h^k, a_h^k\right) &\leq \frac{\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)}{H} + \frac{2HSL}{N'_{kh}\left(s_h^k, a_h^k\right)} + \frac{4H^2SL}{N'_{kh}\left(s_h^k, a_h^k\right)} \\
&\quad + \sqrt{\frac{4L}{N'_{kh}\left(s_h^k, a_h^k\right)}} \left[\left(\sum_{s' \in G_{kh}} P_h\left(s' \mid s_h^k, a_h^k\right) \frac{\tilde{\Delta}_{h+1}^k(s')}{\sqrt{P_h\left(s' \mid s_h^k, a_h^k\right)}} \right) - \frac{\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)}{\sqrt{P_h\left(s_{h+1}^k \mid s_h^k, a_h^k\right)}} \right]
\end{aligned}$$

where $L = \log(S^2AH\pi^2/6\delta')$ *and*

$$G_{kh} \coloneqq \{s' : P_h\left(s' \mid s_h^k, a_h^k\right) N'_{kh}\left(s_h^k, a_h^k\right) \geq 4H^2L\}$$

for all $\mathcal{S} \times \mathcal{A} \times \mathcal{H}$ *and* $K \in \mathbb{N}_1$.

Proof. For completeness, we present proof of this claim and note that the ideas found here were first introduced by Azar et al. (2017).

We upper bound the so-called "correction term", $C\langle \hat{P}_{kh}\left(\cdot \mid s_h^k, a_h^k\right) - P_h\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k \rangle$. Following Azar et al. (2017) and applying Bernstein's inequality to bound the difference between the estimated and actual transitions gives us, with probability $1-\delta$:

$$\begin{aligned}
\gamma_{kh}\left(s_h^k, a_h^k\right) &= \left\langle \left(\hat{P}_{kh} - P_h\right)\left(\cdot \mid s_h^k, a_h^k\right), \tilde{\Delta}_{h+1}^k \right\rangle \\
&\leq 2\sum_{s'}\left(\frac{L}{N'_{kh}\left(s_h^k, a_h^k\right)} + \sqrt{\frac{P_h\left(s' \mid s_h^k, a_h^k\right)L}{N'_{kh}\left(s_h^k, a_h^k\right)}}\right)\tilde{\Delta}_{h+1}^k\left(s'\right) && \text{(Bernstein's Inequality)} \\
&\leq 2\left(\frac{HSL}{N'_{kh}\left(s_h^k, a_h^k\right)} + \sum_{s'}\sqrt{\frac{P_h\left(s' \mid s_h^k, a_h^k\right)L}{N'_{kh}\left(s_h^k, a_h^k\right)}}\tilde{\Delta}_{h+1}^k\left(s'\right)\right) \\
&= 2\left(\frac{HSL}{N'_{kh}\left(s_h^k, a_h^k\right)} + \sum_{s' \notin G_{kh}}\sqrt{\frac{P_h\left(s' \mid s_h^k, a_h^k\right)L}{N'_{kh}\left(s_h^k, a_h^k\right)}}\tilde{\Delta}_{h+1}^k\left(s'\right) + \sum_{s' \in G_{kh}}\sqrt{\frac{P_h\left(s' \mid s_h^k, a_h^k\right)L}{N'_{kh}\left(s_h^k, a_h^k\right)}}\tilde{\Delta}_{h+1}^k\left(s'\right)\right) && (12)
\end{aligned}$$

By definition, $P_h(s'|s,a) < 4H^2L/N'_{kh}(s,a)$ whenever $s' \notin G_{kh}$, which follows simply from rearranging terms in the definition of G_{kh}. Therefore,

$$\sum_{s' \notin G_{kh}}\sqrt{\frac{P_h\left(s' \mid s_h^k, a_h^k\right)L}{N'_{kh}\left(s_h^k, a_h^k\right)}}\tilde{\Delta}_{h+1}^k\left(s'\right) \leq \sum_{s' \notin G_{kh}}\frac{2HL}{N'_{kh}\left(s_h^k, a_h^k\right)}\tilde{\Delta}_{h+1}^k\left(s'\right) \leq \frac{2H^2SL}{N'_{kh}\left(s_h^k, a_h^k\right)}$$

Now, we focus on the $s' \in G_{kh}$.

$$\begin{aligned}
&\sum_{s' \in G_{kh}}\sqrt{\frac{P_h\left(s' \mid s_h^k, a_h^k\right)L}{N'_{kh}\left(s_h^k, a_h^k\right)}}\tilde{\Delta}_{h+1}^k\left(s'\right) \\
&= \sqrt{\frac{L}{P_h\left(s_{h+1}^k \mid s_h^k, a_h^k\right)N'_{kh}\left(s_h^k, a_h^k\right)}}\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) - \sqrt{\frac{L}{P_h\left(s_{h+1}^k \mid s_h^k, a_h^k\right)N'_{kh}\left(s_h^k, a_h^k\right)}}\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right) \\
&\quad + \sum_{s' \in G_{kh}} P_h\left(s' \mid s_h^k, a_h^k\right)\sqrt{\frac{L}{P_h\left(s' \mid s_h^k, a_h^k\right)N'_{kh}\left(s_h^k, a_h^k\right)}} \\
&\leq \frac{\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)}{2H} + \sqrt{\frac{L}{N'_{kh}\left(s_h^k, a_h^k\right)}}\left[\left(\sum_{s' \in G_{kh}} P_h\left(s' \mid s_h^k, a_h^k\right)\frac{\tilde{\Delta}_{h+1}^k\left(s'\right)}{\sqrt{P_h\left(s' \mid s_h^k, a_h^k\right)}}\right) - \frac{\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)}{\sqrt{P_h\left(s_{h+1}^k \mid s_h^k, a_h^k\right)}}\right]
\end{aligned}$$

where the inequality follows from the fact that $s' \in G_{kh}$, implying that $P_h\left(s' \mid s_h^k, a_h^k\right)N'_{kh}\left(s_h^k, a_h^k\right) \geq 4H^2L$. Substituting both of the above into (12) gives:

$$\begin{aligned}
\gamma_{kh}\left(s_h^k, a_h^k\right) &\leq \frac{\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)}{H} + \frac{2HSL}{N'_{kh}}\left(s_h^k, a_h^k\right) + \frac{4H^2SL}{N'_{kh}}\left(s_h^k, a_h^k\right) \\
&\quad + \sqrt{\frac{4L}{N'_{kh}\left(s_h^k, a_h^k\right)}}\left[\left(\sum_{s' \in G_{kh}} P_h\left(s' \mid s_h^k, a_h^k\right)\frac{\tilde{\Delta}_{h+1}^k\left(s'\right)}{\sqrt{P_h\left(s' \mid s_h^k, a_h^k\right)}}\right) - \frac{\tilde{\Delta}_{h+1}^k\left(s_{h+1}^k\right)}{\sqrt{P_h\left(s_{h+1}^k \mid s_h^k, a_h^k\right)}}\right]
\end{aligned}$$

completing the proof. □

B.2 Missing Theoretical Results for Delayed Rewards

In this section, we describe how to use active or lazy updating in the setting where only the rewards return in delay. We assume the rewards are stochastic and their expected values are unknown.

In the setting of delayed rewards, the agent returns the state-action pairs $\{s_h^k, a_h^k\}_{h=1}^H$ at the end of episode k, immediately. However, the rewards $\{r_h^k\}_{h=1}^H$ return with a random delay τ_k. Since it is only the rewards that return in delay, we can

estimate the transitions at the start of each episode, as usual. Thus, we apply active or lazy updating to the estimation of the expected reward function only.

For active updating, this amounts to estimating the expected reward function as soon as new feedback arrives:

$$\hat{r}_{kh}(s,a) = \frac{1}{N'_{kh}(s,a)} \sum_{i=1}^{k-1} r_h^i \mathbb{1}\{s_h^i = s, a_h^i = a, i + \tau_i < k\}$$

For lazy updating, this amounts to waiting until the observed number of rewards for a state-action-step triple have doubled before starting a new epoch. When estimating the expected reward function for j^{th} epoch, the base algorithm will use all the available rewards:

$$\hat{r}_{k_j h}(s,a) = \frac{1}{N'_{k_j h}(s,a)} \sum_{i=1}^{k_j-1} r_h^i \mathbb{1}\{s_h^i = s, a_h^i = a, i + \tau_i < k\}$$

Using Hoeffding's inequality, one can construct confidence sets around the above estimators and derive another estimator that is optimistic, with high probability. We derive the width of the confidence set in the proof below.

Theorem 3. *Let $\mathfrak{R}_K^P$ denote the regret of UCRL2 from estimating the transition densities under immediate feedback. Then, with probability $1-\delta$, the regret of UCRL2 under delayed reward is:*

$$\mathfrak{R}_K \lesssim \mathfrak{R}_K^P + HSA\psi_K^\tau$$

for active updating.

Proof. First, since the rewards are stochastic and their expected values are unknown, we must derive an estimator. Naturally, we use only the observed information to compute the expected value, as it is an unbiased estimator:

$$\hat{r}_{kh}(s,a) = \frac{1}{N'_{kh}(s,a)} \sum_{i=1}^{k-1} r_h^i \mathbb{1}\{s_h^i = s, a_h^i = a, i + \tau_i < k\}$$

Now, assume that the rewards are bounded in $[0,1]$. Using Hoeffding's inequality, we can define an additional failure event to account for the fact that we are estimating the expected reward function:

$$F_k^r = \left\{ \exists\, s,a,h : |\hat{r}_{kh}(s,a) - r_h(s,a)| \geq \sqrt{\frac{6\log(2SAT\pi/6\delta')}{N'_{kh}(s,a)}} := \epsilon_{kh}^r(s,a) \right\}$$

which holds across all episodes with probability $1-\delta'$. Recall, we have a failure event for the transitions that holds with probability $1-\delta'$ too. Thus, we get the following optimistic estimator of the expected reward function:

$$\tilde{r}_{kh}(s,a) = \min\left\{ 1, \hat{r}_{kh}(s,a) + \sqrt{\frac{6\log(2SAT\pi/6\delta')}{N'_{kh}(s,a)}} \right\}$$

which upper bounds the true expected reward function with probability $1-\delta'$ across all episodes. As in the immediate feedback setting, the failure event for the transition densities is:

$$F_k^p = \left\{ \exists\, s,a,h : \|\hat{P}_{kh}(\cdot|s,a) - P_h(\cdot|s,a)\|_1 \geq \sqrt{\frac{6S\log(AT\pi/6\delta')}{N_{kh}(s,a)}} := \epsilon_{kh}^p(s,a) \right\}$$

where

$$\tilde{P}_{kh}(\cdot|s,a) \in \{Q \in \Delta : \|Q - P_h(\cdot|s,a)\| \leq \epsilon_{kh}^p(s,a)\}$$

By optimism, and due to UCRL2 having $C = 1$: with probability $1 - 2\delta'$:

$$\begin{aligned}
\mathfrak{R}_K &= \sum_{k=1}^{K} \Delta_1^k = \sum_{k=1}^{K} V_1^* \left(s_1^k\right) - V_1^{\pi_k} \left(s_1^k\right) \\
&\leq \sum_{k=1}^{K} \tilde{\Delta}_1^k = \sum_{k=1}^{K} \tilde{V}_1^{\pi_k} \left(s_1^k\right) - V_1^{\pi_k} \left(s_1^k\right) \qquad (13) \\
&\leq \sum_{k=1}^{K} \sum_{h=1}^{H} 2H\epsilon_{kh}^p \left(s_h^k, a_h^k\right) + 2\epsilon_{kh}^r \left(s_h^k, a_h^k\right) + \zeta_h^k \left(s_h^k, a_h^k\right) \\
&\leq 2H\sqrt{T \log\left(\frac{K\pi}{6\delta'}\right)} + \sum_{k=1}^{K} \sum_{h=1}^{H} 2H\epsilon_{kh}^p \left(s_h^k, a_h^k\right) + \sum_{k=1}^{K} \sum_{h=1}^{H} 2\epsilon_{kh}^r \left(s_h^k, a_h^k\right) \\
&\leq 2H\sqrt{T \log\left(\frac{K\pi}{6\delta'}\right)} + 2H\sqrt{6S \log\left(AT\pi/6\delta'\right)} \sum_{k=1}^{K} \sum_{h=1}^{H} \frac{1}{\sqrt{N_{kh}\left(s_h^k, a_h^k\right)}} + \sum_{k=1}^{K} \sum_{h=1}^{H} 2\epsilon_{kh}^r \left(s_h^k, a_h^k\right) \\
&\leq 2H\sqrt{T \log\left(\frac{K\pi}{6\delta'}\right)} + 4H\sqrt{6SHSAT \log\left(AT\pi/6\delta'\right)} + \sum_{k=1}^{K} \sum_{h=1}^{H} 2\epsilon_{kh}^r \left(s_h^k, a_h^k\right) \\
&\leq 2H\sqrt{T \log\left(\frac{K\pi}{6\delta'}\right)} + 10H^{3/2}S\sqrt{AT \log\left(AT\pi/6\delta'\right)} + \sum_{k=1}^{K} \sum_{h=1}^{H} 2\epsilon_{kh}^r \left(s_h^k, a_h^k\right) \\
&\leq \mathfrak{R}_K^P + \sum_{k=1}^{K} \sum_{h=1}^{H} 2\epsilon_{kh}^r \left(s_h^k, a_h^k\right) \qquad (14)
\end{aligned}$$

The penultimate inequality follows from Lemma 6. Further, $\mathfrak{R}_K^P = 2H\sqrt{T \log\left(\frac{K\pi}{6\delta'}\right)} + 10H^{3/2}S\sqrt{AT \log\left(AT\pi/6\delta'\right)}$ is the regret of the base algorithm (UCRL2) in an immediate feedback environment with know reward functions. Now, to prove the statements of the corollary, we must bound the summation of the estimation error for the rewards. Doing so is just a matter of applying Lemma 3:

$$\begin{aligned}
\sum_{k=1}^{K} \sum_{h=1}^{H} 2\epsilon_{kh}^r \left(s_h^k, a_h^k\right) &= 2\sum_{k=1}^{K} \sum_{h=1}^{H} \sqrt{\frac{6 \log\left(2SAT\pi/6\delta'\right)}{N'_{kh}\left(s, a\right)}} \\
&\leq 8\sqrt{6HSAT \log\left(SAT\pi/6\delta'\right)} + 6HSA\psi_K^\tau \sqrt{6 \log\left(SAT\pi/6\delta'\right)}
\end{aligned}$$

Substituting the above into Equation (14) and omitting poly-logarithmic factors gives the stated result.

□

C ADDITIONAL EXPERIMENTS

Here, we present additional experimental results for the chain environments with $H = S \in \{5, 10, 20\}$ and $\mathbb{E}[\tau] \in \{100, 300, 500\}$. In all combinations of chain length and expected delay, our updating procedures give better empirical performance, especially for the delay distributions with higher variances. For all expected delays, active updating gives the best performance. However, our experiments indicate that lazy updating with $\alpha = 10$ or 100 is comparable, as one would expect based on the intuition that it is an approximation to active updating that converges in the limit as $\alpha \to \infty$.

C.1 Chain Environment with $H = S = 5$

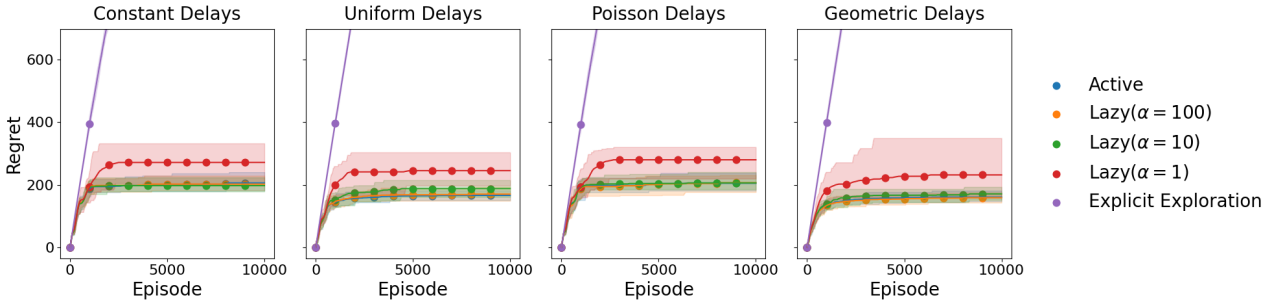

Figure 3: Cumulative Regret $(S = 5, \mathbb{E}[\tau] = 100)$.

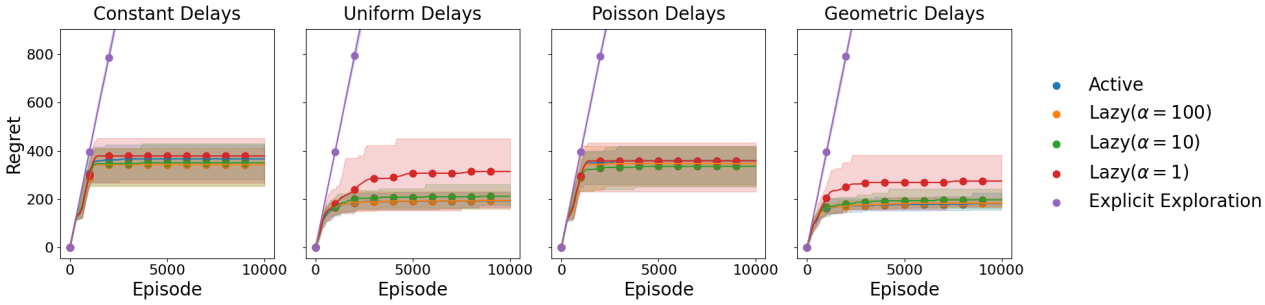

Figure 4: Cumulative Regret $(S = 5, \mathbb{E}[\tau] = 300)$.

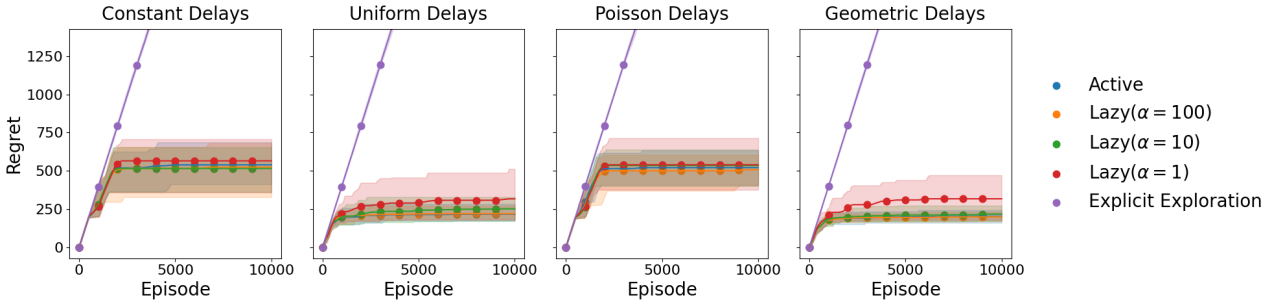

Figure 5: Cumulative Regret $(S = 5, \mathbb{E}[\tau] = 500)$.

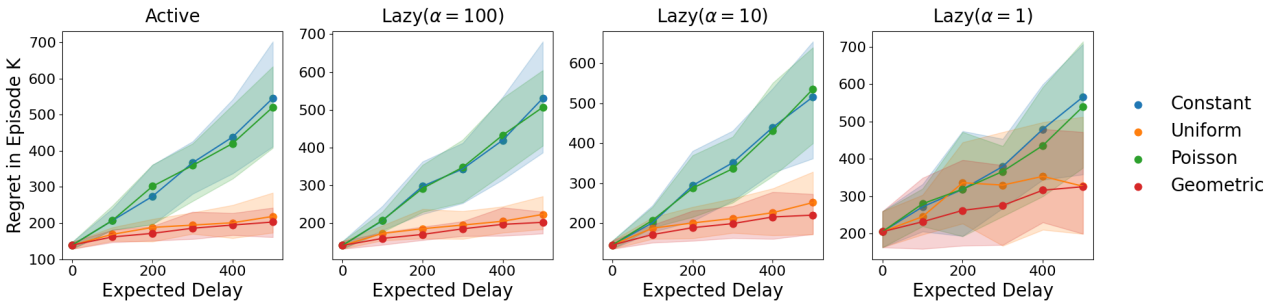

Figure 6: Delay Dependence ($S = 5$)

C.2 Chain Environment with $H = S = 10$

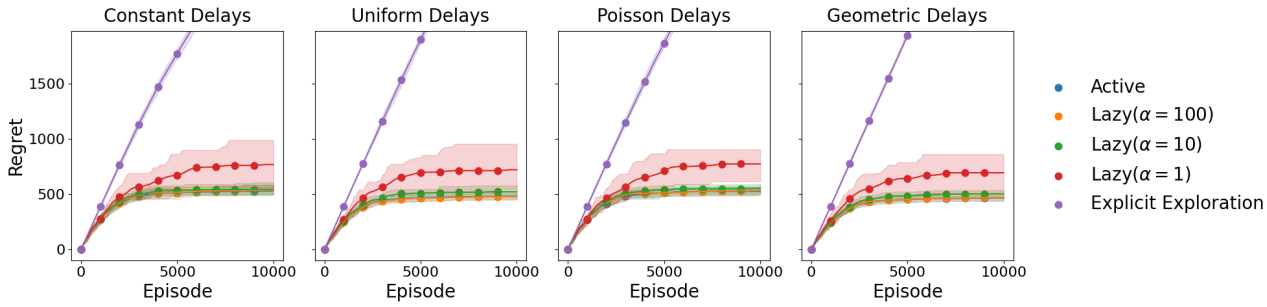

Figure 7: Cumulative Regret ($S = 10, \mathbb{E}[\tau] = 100$).

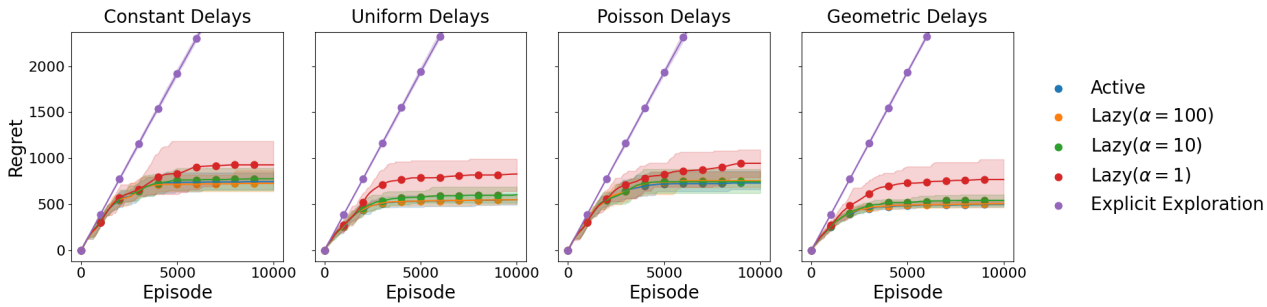

Figure 8: Cumulative Regret ($S = 10, \mathbb{E}[\tau] = 300$).

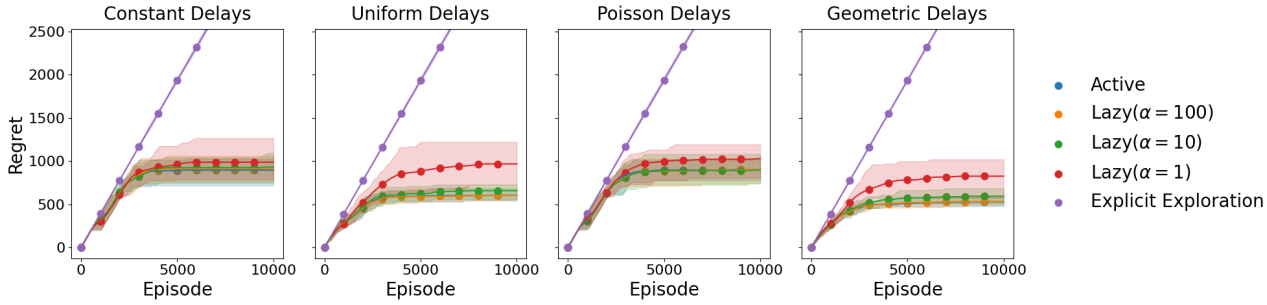

Figure 9: Cumulative Regret ($S = 10, \mathbb{E}[\tau] = 500$).

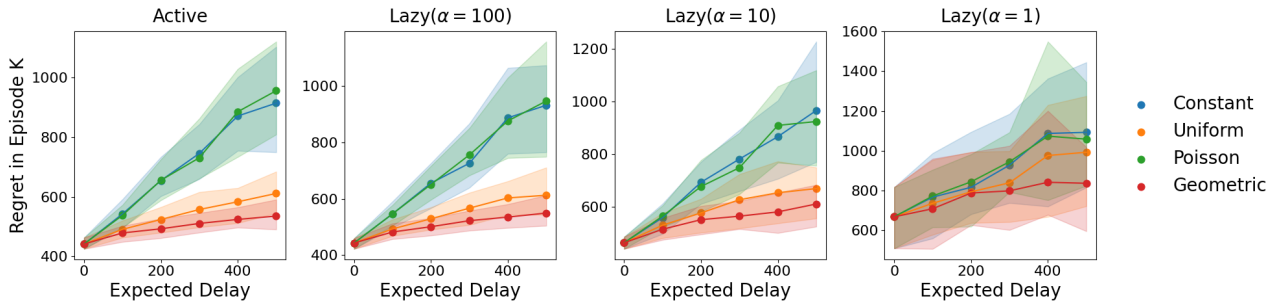

Figure 10: Delay Dependence ($S = 10$)

C.3 Chain Environment with $H = S = 20$

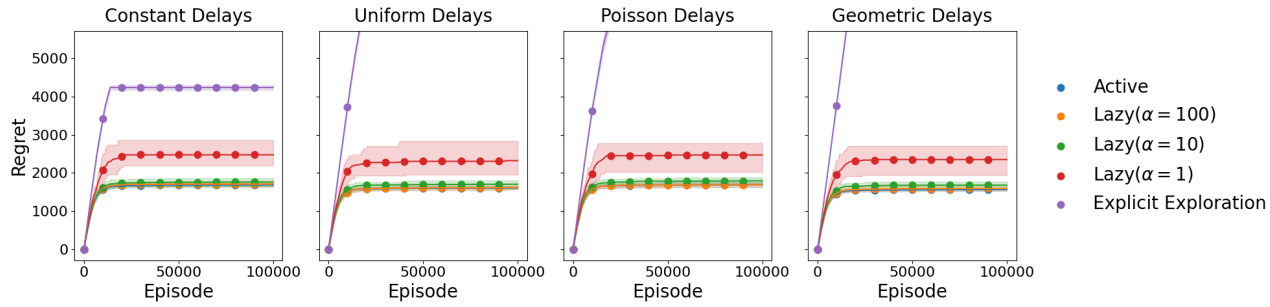

Figure 11: Cumulative Regret ($S = 20, \mathbb{E}[\tau] = 100$).

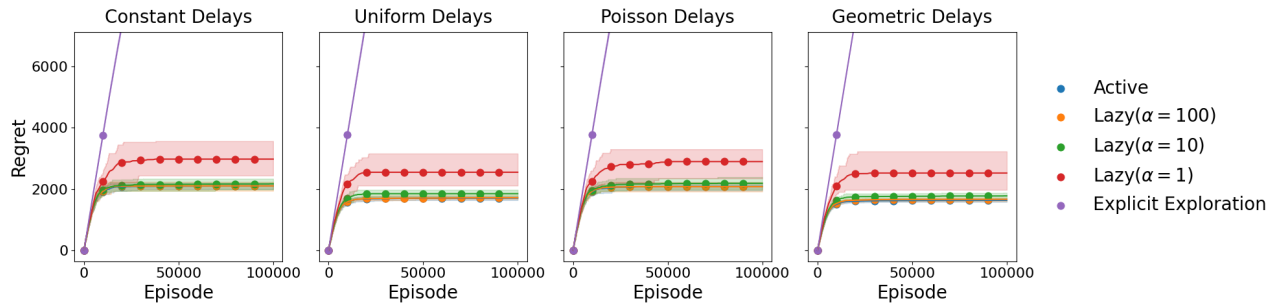

Figure 12: Cumulative Regret ($S = 20, \mathbb{E}[\tau] = 300$).

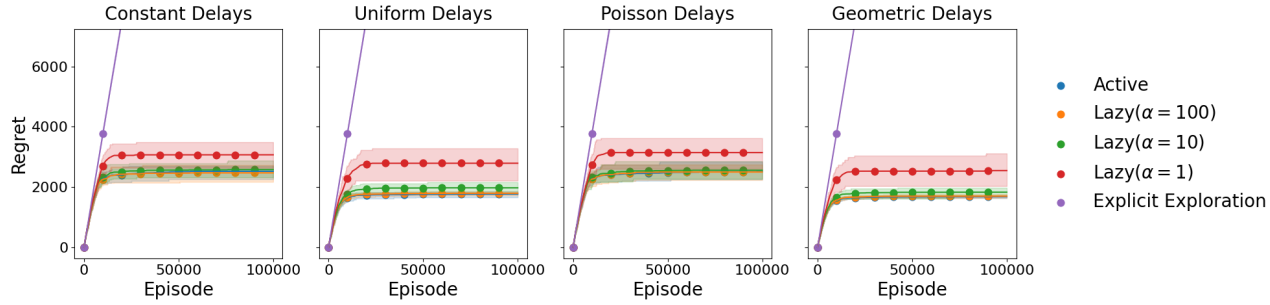

Figure 13: Cumulative Regret ($S = 20, \mathbb{E}[\tau] = 500$).

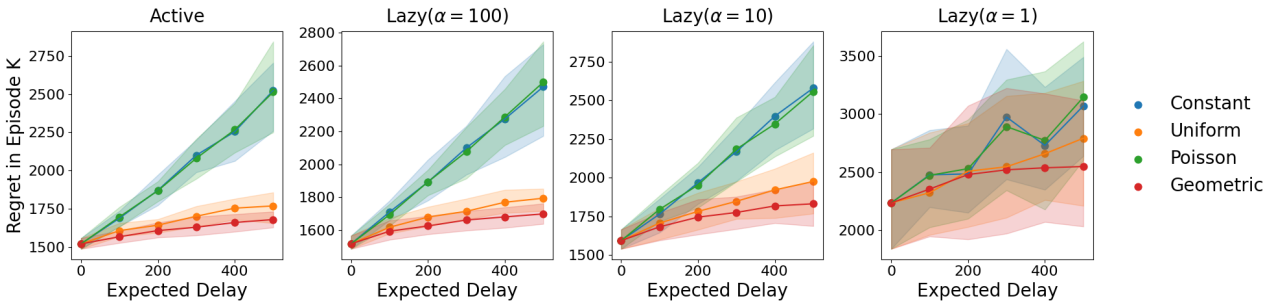

Figure 14: Delay Dependence ($S = 20$)

C.4 Chain Environment with $H = S = 30$

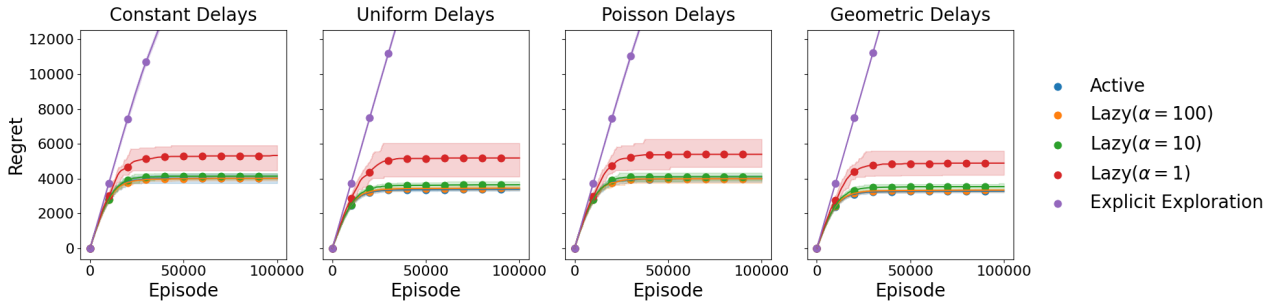

Figure 15: Cumulative Regret ($S = 30, \mathbb{E}[\tau] = 300$).

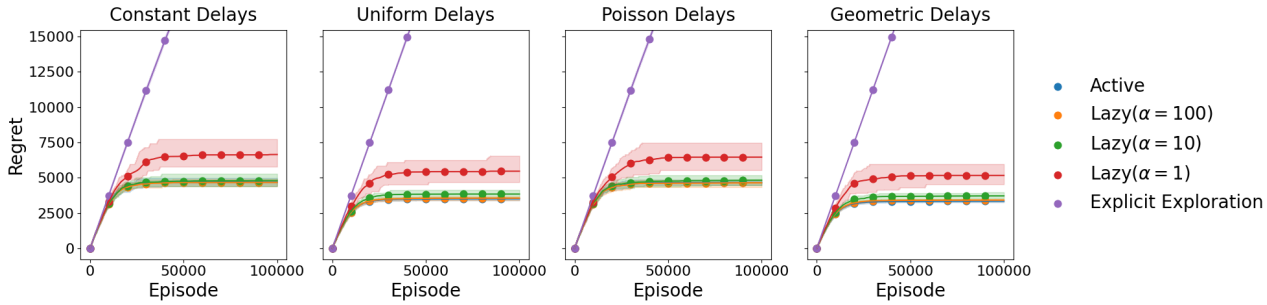

Figure 16: Cumulative Regret ($S = 30, \mathbb{E}[\tau] = 500$).