A Example MDP in Figure 1

This section provides extra details concerning the example MDP in Figure 1. At the end of phase 2, state s_4 has been visited T times each from s_3 and s_1. The visits from s_3 result in an average reward of $g^* + \beta$ while the visits from s_1 result in average reward $g^* - \alpha$, the expected value is therefore:

$$g_4 = \frac{T(g^* + \beta) + T(g^* - \alpha)}{2T} = g^* + \frac{\beta - \alpha}{2}.$$

State s_1, on the other hand, has T transitions each to s_2 and s_4. Its expected value is therefore:

$$g_1 = \frac{T(g^* + \beta) + Tg_4}{2T} = g^* + \frac{3\beta - \alpha}{4}.$$

The total rewards for all $3T$ steps is given by:

$$2T(g^* + \beta) + T(g^* - \alpha) = 3Tg^* + T(2\beta - \alpha).$$

A.1 Additional Details on Experiments

Figure 5 shows the exact MDP used in the experiments. Transitions from s_5, s_6 and s_7 give rewards r_1, r_2 and r_3 respectively. All other transitions give zero rewards. To reproduce the results in Fig. 4, set r_1 to 0.36, r_2 to 1 and r_3 to 0.04. Note that a wide range of parameters can produce similar results.

We run the standard version of UCRL2 as published in [Jaksch et al., 2010]. For OLRM2 we run the exact version as described in this paper. The uncertainty sets are as follows[3]:

- (s_0, a_1): Any distributions over s_0 and s_5
- s_1: Any distributions over s_2 and s_4
- s_4: Any distributions over s_6 and s_7

All other transitions are assumed deterministic. During phase 1, all transitions from s_1 and s_4 go to s_2 and s_6 respectively (solid arrows). In phase 2, all transitions from s_1 and s_4 go to s_4 and s_7 respectively (dashed arrows).

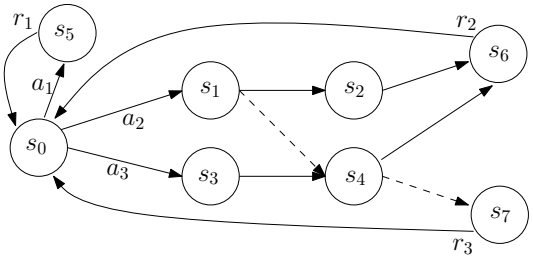

Figure 5: MDP with adversarial transitions.

B The OLRM2 algorithm

The corresponding algorithm for the infinite-horizon average reward case is shown in Figure 6. Section B.1 provides the necessary details to complete the algorithm.

[3]Technically, the MDP would not be communicating if, for instance, s_4 always transitions to s_6, making s_7 unreachable in this case. While it does not affect the results in this example, one can "fix" this by allocating a small minimum probability to all potential next-states.

Input: $\mathcal{S}$, $\mathcal{A}$, δ, and for each (s,a), $\mathcal{U}(s,a)$

1. Initialize the set $F \leftarrow \{\}$.
2. Initialize $k \leftarrow 1$.
3. Compute an optimistic policy $\tilde{\pi}^k$ and obtain its bias h_k, assuming all state-action pairs in F are adversarial (Section B.1).
4. Execute $\tilde{\pi}^k$ until one of the followings happen:
   - The execution count of some state-action (s,a) has been doubled.
   - The executed state-action pair (s,a) fails the stochastic check (Section 5). In this case (s,a) is added to F.
5. Increment k. Go back to step 3.

Figure 6: The OLRM2 algorithm

B.1 Computing an optimistic policy

We adapt the algorithms from [Tewari and Bartlett, 2007] to compute an optimistic minimax policy. The key addition is a stopping condition based on the check for the optimality condition (for unichain MDPs) after every L iterations of the algorithm. Since the convergence rate is geometric, L can simply be chosen to be a small positive integer. The algorithm is given in Figure 7.

C Proofs for Section 4

C.1 Proof of Lemma 1

Proof. We use the following bound from [Weissman et al., 2003] for 1-norm deviation between true distribution p and empirical distribution $\hat{p}$ over S distinct events from n samples:

$$\Pr(\|\hat{p}(\cdot) - p(\cdot)\|_1 \geq \epsilon) \leq (2^S - 2)\exp(-\frac{n\epsilon^2}{2}).$$

For any particular s, a, t and k, the following thus holds with probability at least $1 - \frac{\delta}{2SATk^2}$,

$$\begin{aligned} & \hat{P}_k(\cdot|s,a)\tilde{V}^k_{t+1}(\cdot) - p_{s,a}(\cdot)\tilde{V}^k_{t+1}(\cdot) \\ \leq\ & \|\hat{P}_k(\cdot|s,a) - p_{s,a}(\cdot)\|_1 T \\ \leq\ & T\sqrt{\frac{2}{N_k(s,a)}\log\frac{2^S 2SATk^2}{\delta}} \\ \leq\ & T\sqrt{\frac{2S}{N_k(s,a)}\log\frac{4SATk^2}{\delta}} \end{aligned}$$

Taking union bounds over all states, actions, t and all epochs completes the proof. □

C.2 Proof of Lemma 2

Proof. Consider a check in epoch κ on a transition from (s,a). Let n be the total number of transitions from (s,a) up to and including this transition. Let $s'_1, \ldots, s'_n$ be the next-states of these transitions. Let $k_1, \ldots, k_n$ and $t_1, \ldots, t_n$ be the corresponding epochs and episode stages during which these transitions happened.

Recall that the check fails if

$$\sum_{j=1}^{n} \hat{P}_{k_j}(\cdot|s,a)\tilde{V}^{k_j}_{t_j+1}(\cdot) - \sum_{j=1}^{n} \tilde{V}^{k_j}_{t_j+1}(s'_j) > 5T\sqrt{nS\log\frac{4SAT\tau^2}{\delta}}.$$

Input: $\mathcal{S}$, $\mathcal{A}$, T, δ, F, k, L, and for each (s,a), $\mathcal{U}(s,a)$, $\hat{P}_k(\cdot|s,a)$ and $N_k(s,a)$.

1. Set $\tilde{V}^{(0)}(s) = 0$ for all s.
2. Set $t \leftarrow 0$.
3.
   - Set $\alpha_t \leftarrow \frac{t+1}{t+2}$.
   - For each $(s,a) \in F$, set
   $$\tilde{Q}^{(t+1)}(s,a) = (1-\alpha_t) r(s,a) + \alpha_t \min_{p \in \mathcal{U}(s,a)} p(\cdot)\tilde{V}^{(t)}(\cdot)$$
   and
   $$\tilde{P}_k(\cdot|s,a) \leftarrow \arg\min_{p \in \mathcal{U}(s,a)} p(\cdot)\tilde{V}^{(t)}(\cdot).$$
   - For each $(s,a) \notin F$, set
   $$\tilde{Q}^{(t+1)}(s,a) = (1-\alpha_t) r(s,a) + \alpha_t \max_{\|p - \hat{P}_k(\cdot|s,a)\|_1 \le \sigma(s,a)} p(\cdot)\tilde{V}^{(t)}(\cdot)$$
   where
   $$\sigma(s,a) = \sqrt{\frac{2S}{N_k(s,a)} \log \frac{4SAk^2}{\delta}}.$$
   Set
   $$\tilde{P}_k(\cdot|s,a) \leftarrow \arg\max_{\|p - \hat{P}_k(\cdot|s,a)\|_1 \le \sigma(s,a)} p(\cdot)\tilde{V}^{(t)}(\cdot).$$
   - For each s, set
   $$\tilde{V}^{(t+1)}(s) = \max_a \tilde{Q}^{(t+1)}(s,a)$$
   and
   $$\tilde{\pi}(s) = \arg\max_a \tilde{Q}^{(t+1)}(s,a).$$
   - If $(t+1) \bmod L = 0$, perform the following:
     - Solve the following system of equations for g and $h(\cdot)$, setting $h(0) = 0$.
     $$\forall s \in \mathcal{S}, \quad g + h(s) = r(s, \tilde{\pi}(s)) + \tilde{P}_k(\cdot|s, \tilde{\pi}(s))h(\cdot).$$
     - Check that the following holds for all $s \in F$,
     $$\tilde{\pi}(s) \in \arg\max_a \{r(s,a) + \min_{p \in \mathcal{U}(s,a)} p(\cdot)h(\cdot)\}$$
     and that the following holds for all $s \notin F$,
     $$\tilde{\pi}(s) \in \arg\max_a \{r(s,a) + \max_{\|p - \hat{P}_k(\cdot|s,a)\|_1 \le \sigma(s,a)} p(\cdot)h(\cdot)\}.$$
     If all of the above hold, then stop, output $\tilde{\pi}^k = \tilde{\pi}$ and $h_k = h$.
4. Set $t \leftarrow t+1$. Go to Step 3.

Figure 7: Computing an optimistic minimax algorithm

We show that if (s,a) is stochastic (i.e. $(s,a) \notin \mathcal{F}$) then the probability that this check fails is less than $\frac{\delta}{2\tau^2}$.

Let
$$X_j = p_{s,a}(\cdot)\tilde{V}^{k_j}_{t_j+1}(\cdot) - e_{s'_j}(\cdot)\tilde{V}^{k_j}_{t_j+1}(\cdot)$$
for $j = 1, \ldots, n$, where $e_{s'_j}(\cdot)$ is a (random) indicator vector. Since $\mathbb{E}(X_j | s'_1, \ldots, s'_{j-1}) = 0$, X_j is a martingale difference sequence with $|X_j| \le T$. By Azuma-Hoeffding inequality,
$$\Pr\left(\sum_{j=1}^n X_j \ge \epsilon\right) \le \exp\left(-\frac{\epsilon^2}{2nT^2}\right).$$

Setting $\epsilon = T\sqrt{2n \log \frac{2\tau^2}{\delta}}$, we have that with probability at least $1 - \frac{\delta}{2\tau^2}$,

$$\sum_{j=1}^{n} X_j = \sum_{j=1}^{n} p_{s,a}(\cdot)\tilde{V}_{t_j+1}^{k_j}(\cdot) - \sum_{j=1}^{n} \tilde{V}_{t_j+1}^{k_j}(s'_j) < T\sqrt{2n \log \frac{2\tau^2}{\delta}}. \tag{2}$$

We therefore have

$$\begin{aligned}
& \sum_{j=1}^{n} \hat{P}_{k_j}(\cdot|s,a)\tilde{V}_{t_j+1}^{k_j}(\cdot) - \sum_{j=1}^{n} \tilde{V}_{t_j+1}^{k_j}(s'_j) \\
\leq & \sum_{j=1}^{n} \hat{P}_{k_j}(\cdot|s,a)\tilde{V}_{t_j+1}^{k_j}(\cdot) - \sum_{j=1}^{n} p_{s,a}(\cdot)\tilde{V}_{t_j+1}^{k_j}(\cdot) + T\sqrt{2n \log \frac{2\tau^2}{\delta}} \\
\leq & \left(\sum_{j=1}^{n} T\sqrt{\frac{2S}{N_{k_j}(s,a)} \log \frac{4SATk_j^2}{\delta}} \right) + T\sqrt{2n \log \frac{2\tau^2}{\delta}} \\
\leq & \left(T\sqrt{2S \log \frac{4SAT\kappa^2}{\delta}} \sum_{k=1}^{\kappa} \frac{v_k(s,a)}{\sqrt{N_k(s,a)}} \right) + T\sqrt{2n \log \frac{2\tau^2}{\delta}} \\
\leq & 5T\sqrt{nS \log \frac{4SAT\tau^2}{\delta}}
\end{aligned}$$

where the second inequality is from Lemma 1. In the third inequality we change the index of the summation to over all epochs up to κ, where $v_k(s,a)$ is the total transitions from (s,a) in epoch k. The final inequality uses the fact that $v_k(s,a) \leq N_k(s,a)$ for all k (as imposed by the algorithm), and that $\sum_{k=1}^{\kappa} \frac{v_k(s,a)}{\sqrt{N_k(s,a)}} \leq (\sqrt{2}+1)\sqrt{n}$ under this condition (Lemma 19 from [Jaksch et al., 2010]).

Taking a union bound over all transitions, the total probability that (2) fails to hold is at most

$$\sum_{\tau=1}^{\infty} \frac{\delta}{2\tau^2} = \frac{\delta}{2} \sum_{\tau=1}^{\infty} \frac{1}{\tau^2} \leq \delta.$$

Adding the failure probability of Lemma 1 completes the proof. □

C.3 Proof of Lemma 3

Proof. Since k remains fixed throughout the proof, we omit the superscript k in all the values $\tilde{V}_t^k$ and $\tilde{Q}_t^k$ to reduce clutter.

This is trivially true for $t = T-1$ since

$$\tilde{V}_{T-1}(s) = \max_a r(s,a) = V_{T-1}^*(s)$$

for all s. We prove by induction for $t = (T-2), \ldots, 0$.

Suppose it holds for $\tilde{V}_{t+1}$, i.e. $\tilde{V}_{t+1}(s) \geq V_{t+1}^*(s)$ for all s. During policy computation, there can be four possible cases for each (s,a):

1. $(s,a) \in \mathcal{F}$ and $(s,a) \in F$.
2. $(s,a) \in \mathcal{F}$ and $(s,a) \notin F$.
3. $(s,a) \notin \mathcal{F}$ and $(s,a) \notin F$.
4. $(s,a) \notin \mathcal{F}$ and $(s,a) \in F$.

We will deal with the first 3 cases since the lemma assumes that the last case never happens.

If $(s,a) \in \mathcal{F}$ and has been added to F, then it holds that

$$\tilde{Q}_t(s,a) = r(s,a) + \min_{p \in \mathcal{U}(s,a)} p(\cdot)\tilde{V}_{t+1}(\cdot) \geq r(s,a) + \min_{p \in \mathcal{U}(s,a)} p(\cdot)V^*_{t+1}(\cdot) = Q^*_t(s,a).$$

If (s,a) has not been added to F, then $\tilde{Q}_t(s,a)$ is computed based on $\hat{P}_k(\cdot|s,a)$, which is based on $n = N_k(s,a)$ past transitions of (s,a). Let $s'_1, \ldots, s'_n$ be the next-states of these transitions. Let

$$X_j = p^{(j)}_{s,a}(\cdot)V^*_{t+1}(\cdot) - e_{s'_j}(\cdot)V^*_{t+1}(\cdot)$$

for $j = 1, \ldots, n$, where $e_{s'_j}(\cdot)$ is a random vector with $e_{s'_j}(s') = 1$ if $s'_j = s'$ and zero elsewhere. Note that $p^{(j)}_{s,a} \in \mathcal{U}(s,a)$ for each j and since (s,a) is adversarial $p^{(j)}_{s,a}$ can depend on past transitions $s'_1, \ldots, s'_{j-1}$. However, since $\mathbb{E}(X_j | s'_1, \ldots, s'_{j-1}) = 0$, X_j is a martingale difference sequence with $|X_j| \leq T$. By Azuma-Hoeffding inequality

$$\Pr\left(\sum_{j=1}^n X_j \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2nT^2}\right).$$

Setting $\epsilon = T\sqrt{2n \log \frac{2SATk^2}{\delta}}$, it follows that with failure probability at most $\frac{\delta}{2SATk^2}$,

$$\frac{\sum_{j=1}^n X_j}{n} = \frac{1}{n}\sum_{j=1}^n p^{(j)}_{s,a}(\cdot)V^*_{t+1}(\cdot) - \hat{P}_k(\cdot|s,a)V^*_{t+1}(\cdot) < T\sqrt{\frac{2}{n}\log\frac{2SATk^2}{\delta}}$$

and therefore

$$\begin{aligned}
\tilde{Q}_t(s,a) &= r(s,a) + \hat{P}_k(\cdot|s,a)\tilde{V}_{t+1}(\cdot) + T\sqrt{\frac{2}{n}\log\frac{2SATk^2}{\delta}} \\
&\geq r(s,a) + \hat{P}_k(\cdot|s,a)V^*_{t+1}(\cdot) + T\sqrt{\frac{2}{n}\log\frac{2SATk^2}{\delta}} \\
&\geq r(s,a) + \frac{1}{n}\sum_{j=1}^n p^{(j)}_{s,a}(\cdot)V^*_{t+1}(\cdot) \\
&\geq r(s,a) + \min_{p \in \mathcal{U}(s,a)} p(\cdot)V^*_{t+1}(\cdot) \\
&= Q^*_t(s,a).
\end{aligned}$$

The third case, where $(s,a) \notin \mathcal{F}$, can be proved using similar arguments. In particular, this is simpler since $p_{s,a}$ is fixed for every transition.

We now have

$$\tilde{V}_t(s) = \max_a \tilde{Q}_t(s,a) \geq \max_a Q^*_t(s,a) = V^*_t(s).$$

Taking the union bound over all states, actions, t and k completes the proof. □

C.4 Proof of Theorem 1

Proof. We assume that no state-action pairs $(s,a) \notin \mathcal{F}$ has been added to F. By Lemma 2 this fails with probability at most 2δ.

Assume that episode i is fully within epoch k. Let s^i_t and a^i_t be the state and action taken at stage t in episode i. Let

$$\tilde{\Delta}^i_t = \tilde{V}^k_t(s^i_t) - \sum_{t'=t}^{T-1} r(s^i_{t'}, a^i_{t'}).$$

Then

$$\tilde{\Delta}^i_t = \tilde{V}^k_t(s^i_t) - \sum_{t'=t}^{T-1} r(s^i_{t'}, a^i_{t'})$$

$$
\begin{aligned}
&= \tilde{Q}_t^k(s_t^i, a_t^i) - \sum_{t'=t}^{T-1} r(s_{t'}^i, a_{t'}^i) \\
&= Y_t^i + \tilde{V}_{t+1}^k(s_{t+1}^i) - \sum_{t'=t+1}^{T-1} r(s_{t'}^i, a_{t'}^i) \\
&= Y_t^i + \tilde{\Delta}_{t+1}^i
\end{aligned}
$$

where we have defined

$$
Y_t^i = \begin{cases} \min_{p \in \mathcal{U}(s_t^i, a_t^i)} p(\cdot)\tilde{V}_{t+1}^k(\cdot) - \tilde{V}_{t+1}^k(s_{t+1}^i) & \text{if } (s_t^i, a_t^i) \in F, \\ \hat{P}_k(\cdot|s_t^i, a_t^i)\tilde{V}_{t+1}^k(\cdot) - \tilde{V}_{t+1}^k(s_{t+1}^i) + T\sqrt{\frac{2}{N_k(s_t^i, a_t^i)} \log \frac{2SATk^2}{\delta}} & \text{if } (s_t^i, a_t^i) \notin F. \end{cases}
$$

Since $\tilde{\Delta}_{T-1}^i = 0$, we have that

$$
\tilde{\Delta}_0^i = \sum_{t=0}^{T-2} Y_t^i.
$$

By Lemma 3, with probability at least $1 - \delta$, the regret in episode i is given by

$$
\Delta_i = V_0^*(s_0^i) - \sum_{t=0}^{T-1} r(s_t^i, a_t^i) \leq \tilde{\Delta}_0^i.
$$

Note that the above only holds if there is no change of policy within the episode. Now, after running the algorithm for m episodes, let $\tau = mT$ be the total number of steps executed and κ be the total number of epochs. The total number of new epochs due to doubling of visit counts to state-action pairs can be bounded by $SA \log_2 2\tau$ and the total number of new epochs due to adding a state-action pair to F is at most SA, therefore the total number of epochs is at most $SA \log_2 4\tau$. This means that the total number of episodes with a change of policy is at most $SA \log_2 4\tau$ and the regrets in these episodes can be bounded by $SAT \log_2 4\tau$. For these episodes we define $\tilde{\Delta}_0^i = T$.

The total regret, after running the algorithm for m episodes is therefore

$$
\Delta(m) \leq \sum_{i=1}^{m} \tilde{\Delta}_0^i \leq (SAT \log_2 4\tau) + \sum_{i=1}^{m} \sum_{t=0}^{T-2} Y_t^i. \tag{3}
$$

We now bound the term $\sum_{i=1}^{m} \sum_{t=0}^{T-2} Y_t^i$. Let $n(s, a)$ be the total number of times (s, a) has been executed. Re-write the summation such that it is over state-action pairs:

$$
\sum_{i=1}^{m} \sum_{t=0}^{T-2} Y_t^i = \sum_{s,a} \sum_{j=1}^{n(s,a)} Y_j(s, a).
$$

Fix a state-action pair (s, a). Let $n = n(s, a)$. Let $s'_1, \ldots, s'_n$ be the corresponding next-states in each of the transitions from (s, a). Similarly, let $k_1, \ldots, k_n$ and $t_1, \ldots, t_n$ be the respective epochs and stages when these transitions happen.

Let n_0 be the number of transitions from (s, a) where $(s, a) \notin F$. If (s, a) is never added to F then $n_0 = n$. We have

$$
\begin{aligned}
\sum_{j=1}^{n_0} Y_j(s, a) &= \sum_{j=1}^{n_0} \hat{P}_{k_j}(\cdot|s, a)\tilde{V}_{t_j+1}^{k_j}(\cdot) - \sum_{j=1}^{n_0} \tilde{V}_{t_j+1}^{k_j}(s'_j) + \sum_{j=1}^{n_0} T\sqrt{\frac{2}{n_0} \log \frac{2SATk_j^2}{\delta}} \\
&\leq 5T\sqrt{nS \log \frac{4SAT\tau^2}{\delta}} + \sum_{j=1}^{n_0} T\sqrt{\frac{2}{n_0} \log \frac{2SATk_j^2}{\delta}} \\
&\leq 5T\sqrt{nS \log \frac{4SAT\tau^2}{\delta}} + T\sqrt{2 \log \frac{2SAT\kappa^2}{\delta}} \sum_{k=1}^{\kappa} \frac{v_k(s, a)}{\sqrt{N_k(s, a)}} \\
&\leq 9T\sqrt{nS \log \frac{4SAT\tau^2}{\delta}}
\end{aligned}
$$

where we use the condition for the stochastic check in the first inequality. The second and third inequalities follow the same argument as in Lemma 2.

If (s, a) is ever added to F, then for all subsequent transitions of (s, a),

$$\begin{aligned}\sum_{j=n_0+1}^{n} Y_j(s,a) &= \sum_{j=n_0+1}^{n} \min_{p\in\mathcal{U}(s,a)} p(\cdot)\tilde{V}^{k_j}_{t_j+1}(\cdot) - \tilde{V}^{k_j}_{t_j+1}(s'_j) \\ &\leq \sum_{j=n_0+1}^{n} p^{(j)}_{s,a}(\cdot)\tilde{V}^{k_j}_{t_j+1}(\cdot) - \tilde{V}^{k_j}_{t_j+1}(s'_j) \\ &= \sum_{j=n_0+1}^{n} X_j \\ &< T\sqrt{2n\log\frac{2\tau^2}{\delta}}\end{aligned}$$

where X_j is as defined in Lemma 2 and we apply the Azuma-Hoeffding inequality, with the total probability of failure over all τ at most δ.

Combining the two cases, and taking the sum over all state-action pairs, we have that

$$\sum_{s,a}\sum_{j=1}^{n(s,a)} Y_j(s,a) \leq 11T\sqrt{S\log\frac{4SAT\tau^2}{\delta}}\sum_{s,a}\sqrt{n(s,a)} \leq 11ST\sqrt{A\tau\log\frac{4SAT\tau^2}{\delta}} \quad (4)$$

where the last inequality is by Jensen's inequality.

Combining equations (3) and (4), the total regret is therefore

$$\Delta(m) \leq (SAT\log_2 4\tau) + 11ST\sqrt{A\tau\log\frac{4SAT\tau^2}{\delta}} = \tilde{\mathcal{O}}(ST\sqrt{A\tau}).$$

Adding up all the failure probabilities (by union bound) we get a total failure of 5δ. Run the algorithm with $\frac{\delta}{5}$ and the proof is complete. □

D Analysis (infinite horizon case)

Every time a new policy is computed, a new epoch begins. We use $k = 1, 2, \ldots$ as the index for epochs. The total number of steps from the beginning is indexed by $\tau = 1, 2, \ldots$ and the total number of steps up to the beginning of epoch k is denoted as τ_k.

Lemma 4. *The following holds for all state-action pair $(s, a) \notin \mathcal{F}$ in all epochs $k \geq 1$, with probability at least $1 - \delta$:*

$$\|\hat{P}_k(\cdot|s,a) - p_{s,a}(\cdot)\|_1 \leq \sqrt{\frac{2S}{N_k(s,a)}\log\frac{4SAk^2}{\delta}}.$$

Proof. We use the following bound from [Weissman et al., 2003] for 1-norm deviation between true distribution p and empirical distribution $\hat{p}$ over S distinct events from n samples:

$$\Pr(\|\hat{p}(\cdot) - p(\cdot)\|_1 \geq \epsilon) \leq (2^S - 2)\exp(-\frac{n\epsilon^2}{2}).$$

For any particular s, a and k, the following thus holds with probability at least $1 - \frac{\delta}{2SAk^2}$,

$$\|\hat{P}_k(\cdot|s,a) - p_{s,a}(\cdot)\|_1 \leq \sqrt{\frac{2}{N_k(s,a)}\log\frac{2^S 2SAk^2}{\delta}} \leq \sqrt{\frac{2S}{N_k(s,a)}\log\frac{4SAk^2}{\delta}}$$

Taking union bounds over all states, actions and all epochs completes the proof. □

Lemma 5. *The probability that any state-action pair $(s, a) \notin \mathcal{F}$ gets added into set F while running the algorithm is at most 2δ.*

Proof. Consider a check in epoch κ on a transition from (s,a). Let n be the total number of transitions from (s,a) up to and including this transition. Let $s'_1,\dots,s'_n$ be the next-states of these transitions. Let $k_1,\dots,k_n$ be the corresponding epochs during which these transitions happened.

Recall that the check fails if

$$\sum_{j=1}^{n}\tilde{P}_{k_j}(\cdot|s,a)h_{k_j}(\cdot)-\sum_{j=1}^{n}h_{k_j}(s'_j)>5\tilde{H}\sqrt{nS\log\frac{4SA\tau^2}{\delta}}$$

where $\tilde{H}=\max_{k\in\{k_1,\dots,k_n\}}\left(\max_s h_k(s)-\min_s h_k(s)\right)$.

We show that if (s,a) is stochastic (i.e. $(s,a)\notin\mathcal{F}$) then the probability that this check fails is less than $\frac{\delta}{2\tau^2}$. Note that for (s,a) stochastic, $s'_1,\dots,s'_n$ are independent random variables from the same transition distribution $p_{s,a}$. Consider the function

$$f(s'_1,\dots,s'_n)=\sum_{j=1}^{n}\left(p_{s,a}(\cdot)h_{k_j}(\cdot)-h_{k_j}(s'_j)\right).$$

It is straightforward to show that

$$\max_{s'_1,\dots,s'_n,s''_j}|f(s'_1,\dots,s'_n)-f(s'_1,\dots,s'_{j-1},s''_j,s'_{j+1},\dots,s'_n)|\le\tilde{H}$$

for $1\le j\le n$. We can therefore apply McDiarmid's inequality [McDiarmid, 1989], which gives

$$\Pr\left(f(s'_1,\dots,s'_n)-\mathbb{E}[f(s'_1,\dots,s'_n)]\ge\epsilon\right)\le\exp\left(-\frac{2\epsilon^2}{n\tilde{H}^2}\right).$$

Setting $\epsilon=\tilde{H}\sqrt{\frac{n}{2}\log\frac{2\tau^2}{\delta}}$, and noting that $\mathbb{E}[f(s'_1,\dots,s'_n)]=0$, we have that with probability at least $1-\frac{\delta}{2\tau^2}$,

$$\sum_{j=1}^{n}p_{s,a}(\cdot)h_{k_j}(\cdot)-\sum_{j=1}^{n}h_{k_j}(s'_j)<\tilde{H}\sqrt{\frac{n}{2}\log\frac{2\tau^2}{\delta}}. \tag{5}$$

Let

$$\tilde{U}^{k_j}(\cdot)=h_{k_j}(\cdot)-\frac{\max_{s\in\mathcal{S}}h_{k_j}(s)-\min_{s\in\mathcal{S}}h_{k_j}(s)}{2}.$$

We therefore have

$$\begin{aligned}
&\sum_{j=1}^{n}\tilde{P}_{k_j}(\cdot|s,a)h_{k_j}(\cdot)-\sum_{j=1}^{n}h_{k_j}(s'_j)\\
\le\;&\sum_{j=1}^{n}\tilde{P}_{k_j}(\cdot|s,a)h_{k_j}(\cdot)-\sum_{j=1}^{n}p_{s,a}(\cdot)h_{k_j}(\cdot)+\tilde{H}\sqrt{\frac{n}{2}\log\frac{2\tau^2}{\delta}} &(6)\\
=\;&\sum_{j=1}^{n}\tilde{P}_{k_j}(\cdot|s,a)\tilde{U}^{k_j}(\cdot)-\sum_{j=1}^{n}p_{s,a}(\cdot)\tilde{U}^{k_j}(\cdot)+\tilde{H}\sqrt{\frac{n}{2}\log\frac{2\tau^2}{\delta}}\\
\le\;&\left(\sum_{j=1}^{n}\|\tilde{P}_{k_j}(\cdot|s,a)-p_{s,a}(\cdot)\|_1\frac{\tilde{H}}{2}\right)+\tilde{H}\sqrt{\frac{n}{2}\log\frac{2\tau^2}{\delta}}\\
\le\;&\left(\frac{\tilde{H}}{2}\sum_{j=1}^{n}\left\|\tilde{P}_{k_j}(\cdot|s,a)-\hat{P}_{k_j}(\cdot|s,a)\right\|_1\right)+\left(\frac{\tilde{H}}{2}\sum_{j=1}^{n}\left\|\hat{P}_{k_j}(\cdot|s,a)-p_{s,a}(\cdot)\right\|_1\right)\\
&+\tilde{H}\sqrt{\frac{n}{2}\log\frac{2\tau^2}{\delta}} &(7)\\
\le\;&\tilde{H}\left(\sum_{j=1}^{n}\sqrt{\frac{2S}{N_{k_j}(s,a)}\log\frac{4SAk_j^2}{\delta}}\right)+\tilde{H}\sqrt{\frac{n}{2}\log\frac{2\tau^2}{\delta}}
\end{aligned}$$

$$
\begin{aligned}
&\leq \left(\tilde{H}\sqrt{2S\log\frac{4SA\kappa^2}{\delta}}\sum_{k=1}^{\kappa}\frac{v_k(s,a)}{\sqrt{N_k(s,a)}}\right)+\tilde{H}\sqrt{\frac{n}{2}\log\frac{2\tau^2}{\delta}} \\
&\leq 5\tilde{H}\sqrt{nS\log\frac{4SA\tau^2}{\delta}}
\end{aligned}
$$

where we use (5) in (6). In (7), the first term is bounded by the algorithm when computing the optimistic policy while the second term employs Lemma 4. The last two inequalities follow the same argument as in the proof of Lemma 2.

Taking a union bound over all transitions, the total probability that (2) fails to hold is at most

$$\sum_{\tau=1}^{\infty}\frac{\delta}{2\tau^2}=\frac{\delta}{2}\sum_{\tau=1}^{\infty}\frac{1}{\tau^2}\leq\delta.$$

Adding the failure probability of Lemma 4 completes the proof. □

Lemma 6. *With probability at least $1-\delta$, and assume that no state-action pairs $(s,a)\notin\mathcal{F}$ have been added to F, the following holds for every state $s\in\mathcal{S}$ and every $k\geq 1$:*

$$g^*(s)\leq r(s,\tilde{\pi}^k(s))+\tilde{P}_k(\cdot|s,\tilde{\pi}^k(s))h_k(\cdot)-h_k(s).$$

Proof. Since k remains fixed throughout the proof, we omit the subscript/superscript k from $\tilde{P}$, $\tilde{V}$, $\tilde{Q}$ and $\tilde{\pi}$.

The key is to show that the true minimax MDP is contained in the set of MDPs considered when computing each optimistic policy. This ensures that the optimal gain of the optimistic MDP, $\tilde{g}(s)$, is at least as large as $g^*(s)$ for all $s\in\mathcal{S}$.

During policy computation, there can be four possible cases for each (s,a):

1. $(s,a)\in\mathcal{F}$ and $(s,a)\in F$.
2. $(s,a)\in\mathcal{F}$ and $(s,a)\notin F$.
3. $(s,a)\notin\mathcal{F}$ and $(s,a)\notin F$.
4. $(s,a)\notin\mathcal{F}$ and $(s,a)\in F$.

We only deal with the first 3 cases since the lemma assumes that the last case never happens.

For case 1, the correct set of transition functions is used and therefore the minimax transition is included. For case 2, a transition that is no worse than the minimax transition will always be chosen since it is always chosen optimistically. For case 3, (s,a) is stochastic. By Lemma 4, the true transition function $p_{s,a}$ is included in with probability at least $1-\delta$.

The above implies that $\tilde{g}(s)\geq g^*(s)$ for all s with probability at least $1-\delta$. From equation (1), we therefore have

$$g^*(s)\leq\tilde{g}(s)=r(s,\tilde{\pi}^k(s))+\tilde{P}_k(\cdot|s,\tilde{\pi}^k(s))h_k(\cdot)-h_k(s).$$

□

D.1 Proof of Theorem 2

Proof. We assume that no state-action pairs $(s,a)\notin\mathcal{F}$ has been added to F. By Lemma 5 this fails with probability at most 2δ.

Let s_t and a_t be the state and action taken at step t. Suppose step t is taken in epoch k_t. Consider the regret of this step,

$$
\begin{aligned}
\Delta_t &= g^*(s_t)-r(s_t,a_t) \\
&\leq \tilde{P}_k(\cdot|s_t,a_t)h_{k_t}(\cdot)-h_{k_t}(s_t) \\
&= Y_t+h_{k_t}(s_{t+1})-h_{k_t}(s_t)
\end{aligned}
$$

where the inequality is due to Lemma 6 and we have defined

$$Y_t = \tilde{P}_k(\cdot|s_t, a_t)h_{k_t}(\cdot) - h_{k_t}(s_{t+1}).$$

The total regret in epoch k is then given by

$$\begin{aligned}
\Delta^{(k)} &= \sum_{t=\tau_k+1}^{\tau_{(k+1)}} \Delta_t \\
&\leq \sum_{t=\tau_k+1}^{\tau_{(k+1)}} (Y_t + h_k(s_{t+1}) - h_k(s_t)) \\
&= \left(\sum_{t=\tau_k+1}^{\tau_{(k+1)}} Y_t\right) + h_k(s_{\tau_{(k+1)}+1}) - h_k(s_{\tau_k+1}) \\
&\leq \left(\sum_{t=\tau_k+1}^{\tau_{(k+1)}} Y_t\right) + H
\end{aligned}$$

where in the last inequality we have used the fact that the range of $h_k(\cdot)$ is bounded by the maximal span H.

The total regret, after running the algorithm for τ steps is therefore

$$\Delta(\tau) = \sum_{k=1}^{k_\tau} \Delta^{(k)} \leq \left(\sum_{t=1}^{\tau} Y_t\right) + SAH \log_2 4\tau \tag{8}$$

where we have used the same argument in the proof of Theorem 1 that the total number of epochs is at most $SA \log_2 4\tau$.

We now bound the term $\sum_{t=1}^{\tau} Y_t$. Let $n(s,a)$ be the total number of times (s,a) has been executed. Re-write the summation such that it is over state-action pairs:

$$\sum_{t=1}^{\tau} Y_t = \sum_{s,a} \sum_{j=1}^{n(s,a)} Y_j(s,a).$$

Fix a state-action pair (s,a). Let $n = n(s,a)$. Let $s'_1, \ldots, s'_n$ be the corresponding next-states in each of the transitions from (s,a). Let $k_1, \ldots, k_n$ be the corresponding epochs when these transitions happen.

Suppose (s,a) passes the stochastic check for all n transitions. We have

$$\begin{aligned}
\sum_{j=1}^{n} Y_j(s,a) &= \left(\sum_{j=1}^{n} \tilde{P}_{k_j}(\cdot|s,a)h_{k_j}(\cdot)\right) - \left(\sum_{j=1}^{n} h_{k_j}(s'_j)\right) \\
&\leq 5\tilde{H}\sqrt{nS \log \frac{4SA\tau^2}{\delta}} \\
&\leq 5H\sqrt{nS \log \frac{4SA\tau^2}{\delta}}
\end{aligned}$$

where in the first inequality we use the condition for a successful check.

Note that if (s,a) is ever added to F (say, after n' transitions), then in all subsequent transitions of (s,a),

$$\begin{aligned}
\sum_{j=n'+1}^{n} Y_j(s,a) &= \sum_{j=n'+1}^{n} \min_{p \in \mathcal{U}(s,a)} p(\cdot)h_{k_j}(\cdot) - h_{k_j}(s'_j) \\
&\leq \sum_{j=n'+1}^{n} p^{(j)}_{s,a}(\cdot)h_{k_j}(\cdot) - h_{k_j}(s'_j)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=n'+1}^{n} X_j \\
&\leq H\sqrt{2n \log \frac{2SA\tau^2}{\delta}}
\end{aligned}
$$

where in the last inequality we apply Azuma-Hoeffding inequality to the martingale difference sequence X_j with $|X_j| \leq H$ as in the proof of Theorem 1. The failure probability is at most δ.

We therefore have

$$
\sum_{s,a} \sum_{j=1}^{n(s,a)} Y_j(s,a) \leq 7H\sqrt{S \log \frac{4SA\tau^2}{\delta}} \sum_{s,a} \sqrt{n(s,a)} \leq 7SH\sqrt{A\tau \log \frac{4SA\tau^2}{\delta}} \tag{9}
$$

where the last inequality is by Jensen's inequality.

Combining equations (8) and (9), the total regret is therefore

$$
\Delta(\tau) \leq (SAH \log_2 4\tau) + 7SH\sqrt{A\tau \log \frac{4SA\tau^2}{\delta}} = \tilde{\mathcal{O}}(SH\sqrt{A\tau}).
$$

Adding up all the failure probabilities (by union bound) we get a total failure of 4δ. Run the algorithm with $\frac{\delta}{4}$ and the proof is complete. □