Supplemental material for AAAI-16 paper: Efficient PAC-optimal Exploration in Concurrent, Continuous State MDPs with Delayed Updates

Jason Pazis and Ronald Parr

To prove lemmas 8.1, 8.2, 8.3, and 8.4, and theorem 8.5, we first need to introduce some supporting theory.

12 Bellman error MDPs

This section introduces the concept of Bellman error MDPs, a new analysis tool for RL algorithms. Existing work on bounds based on the Bellman error has focused on one-step max-norm errors (Williams and Baird, 1993). Unfortunately, for the same reason that the maximum expected discounted accumulated reward of an MDP can be much smaller than $\frac{Q_{\max}}{1-\gamma}$, bounds based on the max-norm of the Bellman error can be loose. Bellman error MDPs are much more flexible and intuitive, allowing us to prove stronger bounds with less effort. While this work focuses on PAC optimal exploration, Bellman error MDPs can be used to prove bounds for other online or offline RL algorithms.

Definitions

Let $M$ be an MDP $(S, A, P, R, \gamma)$ with Bellman operator $B^\pi$ for policy $\pi$, and $Q$ an approximate value function for $M$. Additionally, $\forall s, a, \pi$ let $0 \leq Q^\pi(s, a) \leq Q_{\max}$. Let the Bellman error MDP $M_{\epsilon(\pi, Q)}$ be an MDP which differs from $M$ only in its reward function which is defined as $R_{\epsilon(\pi, Q)}(s, a) = Q(s, a) - B^\pi Q(s, a)$ (notice that $Q(s, a)$ and $B^\pi$ are the approximate value function and Bellman operator of the original MDP). We will use $B^\pi_{\epsilon(\pi, Q)}$ to denote $M_{\epsilon(\pi, Q)}$’s Bellman operator under policy $\pi$, and $Q^\pi_{\epsilon(\pi, Q)}$ to denote its value function.

Bellman error MDP theory

Theorem 12.1 is the main theorem of this section. Most results on Bellman error MDPs follow directly from this theorem and basic properties of MDPs.
Theorem 12.1. The return of $\pi$ over $M$ is equal to $Q$ minus the return of $\pi$ over $M_{\epsilon(\pi, Q)}$:

$$Q^\pi(s, a) = Q(s, a) - Q^\pi_{\epsilon(\pi, Q)}(s, a) \quad \forall (s, a) \in (S, A).$$

Proof. We will use induction to prove our claim: If $0$ denotes a value function that is zero everywhere, all we need to prove is that $(B^\pi)^i Q(s, a) = Q(s, a) - (B^\pi_{\epsilon(\pi, Q)})^i 0$ and take the limit as $i \to \infty$.

The base case $B^\pi Q(s, a) = Q(s, a) - B^\pi_{\epsilon(\pi, Q)} 0$ is given by hypothesis. Assuming that $(B^\pi)^i Q(s, a) = Q(s, a) - (B^\pi_{\epsilon(\pi, Q)})^i 0$ holds for $i$, we will prove that it also holds for $i + 1$:

$$(B^\pi)^{i+1} Q(s, a) = B^\pi (B^\pi)^i Q(s, a)$$

$$= \int_{s'} P(s'|s, a) \left( R(s, a, s') + \gamma (B^\pi)^i Q(s', \pi(s')) \right)$$

$$= \int_{s'} P(s'|s, a) \left( R(s, a, s') + \gamma \left( Q(s', \pi(s')) - (B^\pi_{\epsilon(\pi, Q)})^i 0 \right) \right)$$

$$= \int_{s'} P(s'|s, a) \left( R(s, a, s') + \gamma Q(s', \pi(s')) \right) - \int_{s'} P(s'|s, a) \left( \gamma (B^\pi_{\epsilon(\pi, Q)})^i 0 \right)$$

$$= B^\pi Q(s, a) - \int_{s'} P(s'|s, a) \left( \gamma (B^\pi_{\epsilon(\pi, Q)})^i 0 \right)$$

$$= Q(s, a) - R_{\epsilon(\pi)}(s, a) - \int_{s'} P(s'|s, a) \left( \gamma (B^\pi_{\epsilon(\pi, Q)})^i 0 \right)$$

$$= Q(s, a) - (B^\pi_{\epsilon(\pi)})^{i+1} 0$$

If we now take the limit as $i \to \infty$ we have the original claim:

$$\lim_{i \to \infty} (B^\pi)^i Q(s, a) = Q(s, a) - (B^\pi_{\epsilon(\pi, Q)})^i 0 \to$$

$$Q^\pi(s, a) = Q(s, a) - Q^\pi_{\epsilon(\pi, Q)}(s, a)$$

Corollary 12.2 bounds the range of the Bellman error MDP value function, a property that will prove very useful in the analysis of our algorithm. It follows from Theorem 12.1 and the fact that $0 \leq Q^\pi(s, a) \leq Q_{max}$.

Corollary 12.2. Let $0 \leq Q(s, a) \leq Q_{max} \forall (s, a)$. Then:

$$-Q_{max} \leq Q^\pi_{\epsilon(\pi, Q)}(s, a) \leq Q_{max} \forall (s, a, \pi).$$

\footnote{Note that this is true for any policy $\pi$ not just the greedy policy over $Q$.}
Lemma 12.3 below (a consequence of Theorem 12.1), proves that the difference between the optimal and the greedy policy over $Q$ is bounded above by the inverse difference of the value of those policies in their respective Bellman error MDPs.

**Lemma 12.3.** $\forall (s, a, Q)$:

$$V^*(s) - V^{\pi^Q}(s) \leq Q^\pi_{s,\pi^Q,Q}(s, \pi^Q(s)) - Q^\pi_{s,\pi^*,Q}(s, \pi^*(s))$$

**Proof.**

$$Q(s, \pi^*(s)) \leq Q(s, \pi^Q(s)) \Rightarrow$$

$$Q^\pi_{s,\pi^*(s)} + Q^\pi_{\pi^*,Q}(s, \pi^*(s)) \leq Q^\pi_{s,\pi^Q}(s, \pi^Q(s)) + Q^\pi_{\pi^*,Q}(s, \pi^Q(s)) \Rightarrow$$

$$Q^\pi_{s,\pi^*(s)} - Q^\pi_{s,\pi^Q}(s, \pi^Q(s)) \leq Q^\pi_{\pi^*,Q}(s, \pi^Q(s)) - Q^\pi_{\pi^*,Q}(s, \pi^*(s)) \Rightarrow$$

$$V^*(s) - V^{\pi^Q}(s) \leq Q^\pi_{\pi^*,Q}(s, \pi^Q(s)) - Q^\pi_{\pi^*,Q}(s, \pi^*(s))$$

\qed

Lemma 12.4 proves an upper bound on the expected value of the greedy policy in its Bellman error MDP when we do not have a uniform bound on the Bellman error over all state-actions.

**Lemma 12.4.** Let $X_1, \ldots, X_i, \ldots, X_n$ be sets of state-actions where $Q(s, a) - B^s, Q(s, a) \leq \epsilon_i \forall (s, a) \in X_i$, $Q(s, a) - B^\pi^Q, Q(s, a) \leq \epsilon_{\pi^Q} \forall (s, a) \notin \cup_{i=1}^n X_i$, and $\epsilon_{\pi^Q} \leq \epsilon_i \forall i$. Let $p_t^i(s, a, t)$ for $t \in [0, T-1]$ be Bernoulli random variables expressing the probability of taking a state-action $(s_t, a_t)$ at step $t$, for which $(s_t, a_t) \in X_i$, when starting from state-action $(s, a)$ and following $\pi^Q$ thereafter. Then:

$$Q^\pi_{s,\pi^Q,Q}(s, a) \leq \frac{\epsilon_{\pi^Q}}{1 - \gamma} + \epsilon_c,$$

where $\epsilon_c = \sum_{i=1}^n \left( \sum_{t=0}^{T-1} (\gamma^t p_t^i(s, a, t)) (\epsilon_i - \epsilon_{\pi^Q}) \right) + \gamma^T Q_{\text{max}}$.

**Proof.** The expected reward in the Bellman error MDP for step $t \in [0, T-1]$ is bounded above by $\epsilon_{\pi^Q} + \sum_{i=1}^n (p_t^i(s, a, t)) (\epsilon_i - \epsilon_{\pi^Q})$. From Corollary 12.2 the value of any state-action that may be encountered at step $T$ in the Bellman error MDP is bounded above by $Q_{\text{max}}$. From these two facts we have that:

$$Q^\pi_{s,\pi^Q,Q}(s, a) \leq \sum_{t=0}^{T-1} \left( \gamma^t \left( \epsilon_{\pi^Q} + \sum_{i=1}^n (p_t^i(s, a, t)) (\epsilon_i - \epsilon_{\pi^Q}) \right) \right) + \gamma^T Q_{\text{max}}$$

$$\leq \frac{\epsilon_{\pi^Q}}{1 - \gamma} + \sum_{i=1}^n \left( \sum_{t=0}^{T-1} (\gamma^t p_t^i(s, a, t)) (\epsilon_i - \epsilon_{\pi^Q}) \right) + \gamma^T Q_{\text{max}}.$$ 

\qed
Lemmas 12.5 and 12.6 prove bounds on the difference in expected value between the optimal policy and the greedy policy over $Q$ when we do not have a uniform bound on the Bellman error over all state-actions. They avoid the square peg in round hole problem encountered when one tries to analyze exploration algorithms using max-norm bounds.

**Lemma 12.5.** Let $Q(s, a) - B^\pi Q(s, a) \geq -\epsilon_\pi \forall (s, a), X_1, \ldots, X_i, \ldots, X_n$ be sets of state-actions where $Q(s, a) - B^\pi Q(s, a) \leq \epsilon_i \forall (s, a) \in X_i, Q(s, a) - B^\pi Q(s, a) \leq \epsilon_\pi \forall (s, a) \not\in \cup_{i=1}^n X_i$, and $\epsilon_\pi \leq \epsilon_i \forall i$. Let $p_i(s, a, t)$ for $t \in [0, T-1]$ be Bernoulli random variables expressing the probability of taking a state-action $(s_i, a_i)$ at step $t$, for which $(s_i, a_i) \in X_i$, when starting from state-action $(s, a)$ and following $\pi^Q$ thereafter. Then:

$$V^*(s) - V^\pi Q(s) \leq \left(\frac{\epsilon_\pi + \epsilon_\pi^2}{1 - \gamma}\right) + \gamma^T Q_{\text{max}}.$$  

Proof. Since $Q(s, a) - B^\pi Q(s, a) \geq -\epsilon_\pi \forall (s, a), X_1, \ldots, X_i, \ldots, X_n$ be sets of state-actions where $Q(s, a) - B^\pi Q(s, a) \leq \epsilon_i \forall (s, a) \in X_i, Q(s, a) - B^\pi Q(s, a) \leq \epsilon_\pi \forall (s, a) \not\in \cup_{i=1}^n X_i$, and $\epsilon_\pi \leq \epsilon_i \forall i$. Let $T_H = \left\lceil \frac{1}{1 - \gamma} \ln \frac{Q_{\text{max}}}{\epsilon_i} \right\rceil$ and define $H = \{1, 2, 4, \ldots, 2^i\}$ where $i$ is the largest integer such that $2^i \leq T_H$. Define $p_{h,i}(s, a)$ for $h \in [0, T_H - 1]$ to be Bernoulli random variables expressing the probability of encountering exactly $h$ state-actions for which $(s, a) \in X_i$ when starting from state-action $(s, a)$ and following $\pi^Q$ thereafter for a total of $\min\{T, T_H\}$ steps. Finally let $p_{h,i}(s, a) = \sum_{m=h}^{2^i-1} p_{m,i}(s, a)$. Then:

$$V^*(s) - V^\pi Q(s) \leq \left(\frac{\epsilon_\pi + \epsilon_\pi^2}{1 - \gamma}\right) + \epsilon + \epsilon_\pi,$$

where $\epsilon_\pi = 2 \sum_{i=1}^n \sum_{h \in H} (h p_{h,i}(s, a)) (\epsilon_i - \epsilon_\pi) + \gamma T Q_{\text{max}}$.

Proof. Let $p_i(s, a, t)$ for $t \in [0, T-1]$ be Bernoulli random variables expressing the probability of taking a state-action $(s_i, a_i)$ at step $t$, for which $(s_i, a_i) \in X_i$, when starting from state-action $(s, a)$ and following $\pi^Q$ thereafter. We have that $\forall i$:

$$\sum_{i=0}^{T-1} (\gamma^i p_i(s, a, t)) \leq \sum_{i=0}^{T-1} p_i(s, a, t) = \sum_{h=1}^T (h p_{h,i}(s, a)) \leq 2 \sum_{h \in H} (h p_{h,i}(s, a)) .$$

We also have that:

$$\gamma^{\min\{T, T_H\}} Q_{\text{max}} \leq \gamma T T_H Q_{\text{max}} + \gamma T Q_{\text{max}} \leq \epsilon + \gamma^T Q_{\text{max}},$$

and the result follows from Lemma 12.5. $\square$
Using Bellman error MDPs

An important consequence of Lemma 12.3 is that in order to get a bound on the difference in performance between the optimal policy and the greedy policy over a value function, we only need a lower bound on the performance of an optimal policy and an upper bound on the performance of the greedy policy in their respective Bellman error MDPs.

The algorithm presented in this paper relies on an optimistic approximation scheme. Because of that, we will be able to get a good bound on the performance of the optimal policy in its Bellman error MDP from the start, while our bound on the performance of the greedy policy will improve as more samples are gathered.

13 Useful lemmas

Lemma 13.1. Let $t_i$ for $i = 0 \rightarrow l$ be the outcomes of independent (but not necessarily identically distributed) random variables in $\{0, 1\}$, with $P(t_i = 1) \geq p_i$. If $\frac{2}{m} \ln \frac{1}{\delta} < 1$ and:

$$\sum_{i=0}^{l} p_i \geq \frac{m}{1 - \sqrt{\frac{2}{m} \ln \frac{1}{\delta}}}$$

then $\sum_{i=0}^{l} t_i \geq m$ with probability at least $1 - \delta$.

Proof. Let $\sum_{i=0}^{l} p_i = \mu$. From Chernoff’s bound we have that the probability that $\sum_{i=0}^{l} t_i$ is less than $(1 - \epsilon)\mu$ for $0 < \epsilon < 1$ is upper bounded by:

$$\delta = P \left( \sum_{i=0}^{l} t_i \leq (1 - \epsilon)\mu \right) \leq e^{-\frac{\epsilon^2 \mu}{2}}.$$  \hspace{1cm} (1)

Setting $(1 - \epsilon)\mu = m$ we have that $\mu = \frac{m}{1 - \epsilon}$. Substituting into equation 1 above and taking logarithms we have:

$$\ln \delta \leq -\frac{\epsilon^2 \frac{m}{1 - \epsilon}}{2} \Rightarrow$$

$$\frac{\epsilon^2}{1 - \epsilon} \leq \frac{2}{m} \ln \frac{1}{\delta} \Rightarrow$$

$$\epsilon^2 < \frac{2}{m} \ln \frac{1}{\delta} \Rightarrow$$

$$\epsilon < \sqrt{\frac{2}{m} \ln \frac{1}{\delta}}.$$ 

Substituting into $\mu = \frac{m}{1 - \epsilon}$ concludes our proof. \qed
Lemma 13.1 is an improvement over Li’s lemma 56 (Li, 2009) in two ways: While Li’s lemma requires that $\mu > 2m$ for all $\delta < 1$, for realistically large values of $m$ lemma 13.1 yields values much closer to $m$. As an example, for $\delta = 10^{-9}$ and $k = 10^3, 10^6$ and $10^9$ our bound yields $1.2556m, 1.0065m$ and $1.0002m$ respectively. More importantly, our bound allows for the success probability of each trial to be different, which will allow us to prove TCE PAC bounds, rather than just bounds on the number of significantly suboptimal steps.

Lemma 13.2. $\tilde{B}$ is a $\gamma$-contraction in maximum norm.

Proof. Suppose $\|Q_1 - Q_2\|_{\infty} = \epsilon$. For any $(s, a, M)$ we have:

$$\tilde{B}Q_1(s, a, M) = \min \{ Q_{\max}, F(Q_1, N(U, s, a, M)) + d(s, a, M, N(U, s, a, M), d_{\text{known}}) \}$$

$$\leq \min \{ Q_{\max}, F(Q_2, N(U, s, a, M)) + \gamma \epsilon + d(s, a, M, N(U, s, a, M), d_{\text{known}}) \}$$

$$\leq \gamma \epsilon + \min \{ Q_{\max}, F(Q_2, N(U, s, a, M)) + d(s, a, M, N(U, s, a, M), d_{\text{known}}) \}$$

$$= \gamma \epsilon + \tilde{B}Q_2(s, a, M)$$

$$\Rightarrow \tilde{B}Q_1(s, a, M) \leq \gamma \epsilon + \tilde{B}Q_2(s, a, M).$$

Similarly we have that $\tilde{B}Q_2(s, a, M) \leq \gamma \epsilon + \tilde{B}Q_1(s, a, M)$ which completes our proof. \qed

Versions of lemma 13.3 below have appeared in the literature before. It is included here for completeness:

Lemma 13.3. Let $\hat{B}$ be a $\gamma$-contraction with fixed point $\hat{Q}$, and $Q$ the output of

$$\frac{\ln \epsilon}{\ln (Q_{\max} - Q_{\min})}$$

iterations$^2$ of value iteration using $\hat{B}$. Then if $Q_{\min} \leq \hat{Q}(s, a, M) \leq Q_{\max}$ and $Q_{\min} \leq Q_0(s, a, M) \leq Q_{\max} \forall (s, a, M)$, where $Q_0(s, a, M)$ is the initial value for $(s, a, M)$:

$$-\epsilon \leq Q(s, a, M) - \hat{B}Q(s, a, M) \leq \epsilon \forall (s, a, M).$$

Proof. If $\gamma = 0 \tilde{B}Q_0(s, a, M) = \hat{Q}(s, a, M) \forall (s, a, M)$. Otherwise, we have that:

$$||Q - Q_0||_{\infty} \leq Q_{\max} - Q_{\min}.$$

$^2$Note that $\frac{\ln (Q_{\max} - Q_{\min})}{\ln \gamma} \leq \frac{1}{1 - \gamma} \ln \frac{Q_{\max} - Q_{\min}}{\epsilon}$. 

6
Let $Q_n$ be the value function at the $n$-th step of value iteration. Since $\hat{B}$ is a $\gamma$-contraction with fixed point $\hat{Q}$, for $n \geq 1$:

$$||\hat{Q} - \hat{B}Q_n||_\infty = ||\hat{B}\hat{Q} - \hat{B}Q_n||_\infty$$

$$\leq \gamma||\hat{Q} - Q_{n-1}||_\infty$$

$$= \gamma||\hat{Q} - BQ_{n-2}||_\infty$$

$$\leq \gamma^2||\hat{Q} - Q_{n-2}||_\infty$$

$$\ldots$$

$$\leq \gamma^n||\hat{Q} - Q_0||_\infty = \Rightarrow$$

$$||\hat{Q} - \hat{B}Q_n||_\infty \leq \gamma^n(Q_{\text{max}} - Q_{\text{min}}).$$

We want

$$\epsilon \leq \gamma^n(Q_{\text{max}} - Q_{\text{min}}),$$

which after redistributing and taking the $\gamma$ logarithm becomes:

$$n \leq \log_\gamma \frac{\epsilon}{Q_{\text{max}} - Q_{\text{min}}}$$

$$= \frac{\ln \epsilon}{\ln (Q_{\text{max}} - Q_{\text{min}})} / \ln \gamma.$$
Proof. Let $Y$ be the set of $k_{a}(s, a, M)$ samples used by $F^{\pi}(\tilde{Q}, u(s, a, M))$ at $(s, a, M)$ and define $f(x_1, \ldots, x_{k_{a}}) = F^{\pi}(\tilde{Q}, u(s, a, M))$, where $x_1, \ldots, x_{k_{a}}$ are realizations of independent (from the Markov property) variables whose outcomes are possible next states, one for each state-action-MDP in $Y$. The outcomes of the variables (which is where the Markov property ensures independence) are the next states the transitions lead to, not the state-action-MDP triples the samples originate from. The state-action-MDP triples the samples originate from are fixed with respect to $f$, and no assumptions are made about their distribution. Since $\tilde{U}$ is fixed, $\tilde{Q}$ is fixed with respect to $f$ (we are examining the effects of a single application of $F^{\pi}(\tilde{Q}, u(s, a, M))$ to the fixed function $\tilde{Q}$, while varying the next-states that samples in $u(s, a, M)$ land on).

If $k_{a}(s, a, M) > 0$ then $\forall i \in [1, k_{a}(s, a, M)]$:

$$
\sup_{x_1, \ldots, x_k, \hat{x}_i} |f(x_1, \ldots, x_k) - f(x_1, \ldots, x_{i-1}, \hat{x}_i, x_{i+1} \ldots x_{k_{a}(s, a, M)})| = c_i \leq \frac{\hat{Q}_{\text{max}}}{k_{a}(s, a, M)},
$$

and

$$
\sum_{i=1}^{k_{a}(s, a, M)} (c_i)^2 \leq k_{a}(s, a, M) \frac{\hat{Q}_{\text{max}}^2}{k_{a}(s, a, M)^2} = \frac{\hat{Q}_{\text{max}}^2}{k_{a}(s, a, M)}.
$$

From McDiarmid’s inequality we have:

$$
P\left(F^{\pi}(\tilde{Q}, u(s, a, M)) - \mathbb{E}[F^{\pi}(\tilde{Q}, u(s, a, M))] \leq -\epsilon_0\right)
\leq P\left(f(x_1, \ldots, x_{k_{a}}) - \mathbb{E}[f(x_1, \ldots, x_{k_{a}})] \leq -\epsilon_0\right)
\leq e^{-\frac{2\epsilon_0^2}{\sum_{i=1}^{k_{a}(s, a, M)} (c_i)^2}}
\leq e^{-\frac{2k_{a}(s, a, M) \epsilon_0^2}{\hat{Q}_{\text{max}}^2}}.
$$

and:

$$
P\left(F^{\pi}(\tilde{Q}, u(s, a, M)) - \mathbb{E}[F^{\pi}(\tilde{Q}, u(s, a, M))] \geq \epsilon_0\right)
\leq P\left(f(x_1, \ldots, x_{k_{a}}) - \mathbb{E}[f(x_1, \ldots, x_{k_{a}})] \geq \epsilon_0\right)
\leq e^{-\frac{2\epsilon_0^2}{\sum_{i=1}^{k_{a}(s, a, M)} (c_i)^2}}
\leq e^{-\frac{2k_{a}(s, a, M) \epsilon_0^2}{\hat{Q}_{\text{max}}^2}}.
$$

If $k_{a}(s, a, M) = 0$ then $e^{-\frac{2k_{a}(s, a, M) \epsilon_0^2}{\hat{Q}_{\text{max}}^2}} = 1$ and the above is trivially true.
From Definition 5.13 we have that:

\[ B^\pi Q(s, a, M) - \varepsilon_c \leq E \left[ F^\pi (Q, u(s, a, M)) \right] - \frac{\varepsilon_b}{\sqrt{k_a(s, a, M)}} \]

and

\[ E \left[ F^\pi Q (Q, u(s, a, M)) \right] \leq B^\pi Q(s, a, M) + \varepsilon_c + \frac{\varepsilon_b}{\sqrt{k_a(s, a, M)}}. \]

where the expectations are over the next-states that samples in \( u(s, a, M) \) used by \( F^\pi \) and \( F^\pi Q \) respectively land on.

Substituting \( \frac{\varepsilon_b}{\sqrt{k_a(s, a, M)}} \) for \( \varepsilon_0 \) above we have that:

\[
P \left( F^\pi (Q, u(s, a, M)) - B^\pi Q(s, a, M) \leq -\varepsilon_c \right)
\leq P \left( F^\pi (Q, u(s, a, M)) - E \left[ F^\pi (Q, u(s, a, M)) \right] \leq -\frac{\varepsilon_b}{\sqrt{k_a(s, a, M)}} \right)
\leq e^{-2 \left( \frac{\varepsilon_b}{\sqrt{k_a(s, a, M)}} \right)^2 k_a(s, a, M)}
\leq e^{-2 (\text{max}_{\text{next-state}} \left[ 4 \left[ 1 + \log_2 k \right] N_{S,A,M}(d_{\text{known}})^2 \right] k_a(s, a, M))}
\]

and:

\[
P \left( F^\pi Q (Q, u(s, a, M)) - B^\pi Q(s, a, M) \geq \varepsilon_c + \frac{\varepsilon_b}{\sqrt{k_a(s, a, M)}} \right)
\leq P \left( F^\pi Q (Q, u(s, a, M)) - E \left[ F^\pi Q (Q, u(s, a, M)) \right] \geq \frac{\varepsilon_b}{\sqrt{k_a(s, a, M)}} \right)
\leq e^{-2 \left( \frac{\varepsilon_b}{\sqrt{k_a(s, a, M)}} \right)^2 k_a(s, a, M)}
\leq e^{-2 (\text{max}_{\text{next-state}} \left[ 4 \left[ 1 + \log_2 k \right] N_{S,A,M}(d_{\text{known}})^2 \right] k_a(s, a, M))}
\]

Given a bound on the probability that an individual approximation unit has Bellman error of unacceptably high magnitude, lemma 13.5 uses the union
bound to bound the probability that there exists at least one approximation unit in \( \tilde{U} \) for some published \( \tilde{U} \) with Bellman error of unacceptably high magnitude during a non-delay step.

**Lemma 13.5.** If \( \epsilon_b = \hat{Q}_{\text{max}} \sqrt{\ln \frac{4(1 + \log_2 k)N_{\text{RAM}(d_{\text{known}})}}{2}} \), the probability that for any published \( \tilde{U} \) there exists at least one \( u(s, a, M) \in \tilde{U} \) such that:

\[
F^\pi_s(Q_{\tilde{U}}, u(s, a, M)) - B^\pi_s Q_{\tilde{U}}(s, a, M) \leq -\epsilon_c
\]  
\hspace{1cm} (2)

or:

\[
F^{\pi_Q}_{\tilde{U}}(Q_{\tilde{U}}, u(s, a, M)) - B^{\pi_Q}_{\tilde{U}} Q_{\tilde{U}}(s, a, M) \geq \epsilon_c + \frac{c \epsilon_b}{\sqrt{k_a(s, a, M)}}
\]  
\hspace{1cm} (3)

during a non-delay step, is upper bounded by \( \frac{\delta}{2} \).

**Proof.** When \( \tilde{U} \) is the empty set no approximation units exist, so the above is trivially true. Otherwise, at most \( r \log_2 kN_{\text{SAM}(d_{\text{known}})} \) distinct non-empty \( \tilde{U} \) exist during non-delay steps. Thus, there are at most \( 2^{r \log_2 kN_{\text{SAM}(d_{\text{known}})}} \) ways for at least one of the at most \( N_{\text{SAM}(d_{\text{known}})} \) approximation units to fail at least once during non-delay steps \( (1 + \log_2 kN_{\text{SAM}(d_{\text{known}})}) \) ways each for equation 2 or equation 3 to be true at least once), each with a probability at most \( \frac{\epsilon_b}{2} \). From the union bound, we have that the probability that for any published \( \tilde{U} \) there exists at least one \( u(s, a, M) \in \tilde{U} \) such that equation 2 or 3 is true during a non-delay step, is upper bounded by \( \frac{\delta}{2} \). \( \Box \)

Based on lemma 13.5 we can now bound the probability that any \( (s, a, M) \) will have Bellman error of unacceptably high magnitude during a non-delay step:

**Lemma 13.6.** If \( \epsilon_b = \hat{Q}_{\text{max}} \sqrt{\ln \frac{4(1 + \log_2 k)N_{\text{RAM}(d_{\text{known}})}}{2}} \) then:

\[
Q_{\tilde{U}}(s, a, M) - B^\pi_s Q_{\tilde{U}}(s, a, M) > -\epsilon_a - 2\epsilon_c
\]

and:

\[
Q_{\tilde{U}}(s, a, M) - B^{\pi_Q}_{\tilde{U}} Q_{\tilde{U}}(s, a, M) < \epsilon_a + 2\epsilon_c + \frac{\epsilon_b}{\sqrt{k_a(s, a, M)}} + 2d(s, a, M, N(\tilde{U}, s, a, M), d_{\text{known}})
\]

\( \forall (s, a, M, \tilde{U}) \) simultaneously with probability \( 1 - \frac{\delta}{2} \).

**Proof.** When \( \tilde{U} \) is the empty set, \( Q_{\tilde{U}}(s, a, M) = Q_{\text{max}} \). Since \( B^\pi_s Q_{\tilde{U}}(s, a, M) \leq Q_{\text{max}}, Q_{\tilde{U}}(s, a, M) - B^\pi_s Q_{\tilde{U}}(s, a, M) \geq -\epsilon_a - 2\epsilon_c \). Otherwise, \( \forall (s, a, M, \tilde{U}) \) with probability \( 1 - \frac{\delta}{2} \):

\[
B^\pi_s Q_{\tilde{U}}(s, a, M) = \min \{ Q_{\text{max}}, B^\pi_s Q_{\tilde{U}}(s, a, M) \}
\]
\[ \leq \min \left\{ Q_{\text{max}}, B^{\pi_b} Q_{\bar{U}}(\mathcal{U}, s, a, M) \right\} \]
\[ + d(s, a, M, \mathcal{U}, s, a, M, d_{\text{known}}) + \epsilon_c \}
\[ < \min \left\{ \left\{ \max, F^{\pi_b} Q_{\bar{U}}, u(\mathcal{U}, s, a, M) \right\} + \epsilon_c + \right. \]
\[ \left. d(s, a, M, \mathcal{U}, s, a, M, d_{\text{known}}) + \epsilon_c \right\} \]
\[ \leq B^{\pi_b} Q_{\bar{U}}(s, a, M) + 2\epsilon_c \]
\[ \leq B Q_{\bar{U}}(s, a, M) + 2\epsilon_c + 0 \]
\[ \leq Q_{\bar{U}}(s, a, M) + \epsilon_a + 2\epsilon_c. \]

In step 1 we used the fact that \( B^{\pi_b} Q_{\bar{U}}(s, a, M) \leq Q_{\text{max}} \), in step 2 we used Definition 5.13, in step 3 we used lemma 13.5, in step 4 we used Definition 5.11, in step 5 we used the fact that \( B Q_{\bar{U}}(s, a, M) \geq B^{\pi_b} Q_{\bar{U}}(s, a, M) \), and in step 6 we used the fact that \( -\epsilon_a \leq Q_{\bar{U}}(s, a, M) - B Q_{\bar{U}}(s, a, M) \).

When no approximation units containing active samples exist, \( d(s, a, M, \mathcal{U}, s, a, M) \) is not defined. Otherwise, \( \forall (s, a, M, U) \) with probability \( 1 - \frac{11}{2} \).

\[ B^{\pi_{Q_{\bar{U}}}} Q_{\bar{U}}(s, a, M) \]
\[ = \min \left\{ \min, B^{\pi_{Q_{\bar{U}}}} Q_{\bar{U}}(s, a, M) \right\} \]
\[ > \min \left\{ \min, B^{\pi_{Q_{\bar{U}}}} Q_{\bar{U}}(\mathcal{U}, s, a, M) \right\} - d(s, a, M, \mathcal{U}, s, a, M, d_{\text{known}}) - \epsilon_c \}
\[ > \min \left\{ \min, F^{\pi_{Q_{\bar{U}}}} Q_{\bar{U}}, u(\mathcal{U}, s, a, M) \right\} - \epsilon_c - 2 \frac{\epsilon_b}{\sqrt{k_a(s, a, M)}} - \right. \]
\[ \left. \epsilon_c - d(s, a, M, \mathcal{U}, s, a, M, d_{\text{known}}) \right\} \]
\[ \leq \min \left\{ \min, F^{\pi_{Q_{\bar{U}}}} Q_{\bar{U}}, u(\mathcal{U}, s, a, M) \right\} + d(s, a, M, \mathcal{U}, s, a, M, d_{\text{known}}) \}
\[ - 2\epsilon_c - 2 \frac{\epsilon_b}{\sqrt{k_a(s, a, M)}} - 2d(s, a, M, \mathcal{U}, s, a, M, d_{\text{known}}) \]
\[ \geq B^{\pi_{Q_{\bar{U}}}} Q_{\bar{U}}(s, a, M) - 2\epsilon_c - 2 \frac{\epsilon_b}{\sqrt{k_a(s, a, M)}} - 2d(s, a, M, \mathcal{U}, s, a, M, d_{\text{known}}) \]
\[ = B Q_{\bar{U}}(s, a, M) - 2\epsilon_c - 2 \frac{\epsilon_b}{\sqrt{k_a(s, a, M)}} - 2d(s, a, M, \mathcal{U}, s, a, M, d_{\text{known}}) \]
\[ \geq Q_{\bar{U}}(s, a, M) - \epsilon_a - 2\epsilon_c - 2 \frac{\epsilon_b}{\sqrt{k_a(s, a, M)}} - 2d(s, a, M, \mathcal{U}, s, a, M, d_{\text{known}}). \]

In step 1 we used the fact that \( B^{\pi_{Q_{\bar{U}}}} Q_{\bar{U}}(s, a, M) \leq Q_{\text{max}} \), in step 2 we used Definition 5.13, in step 3 we used lemma 13.5, in step 5 we used Definition 5.11, in step 6 we used the fact that \( B^{\pi_{Q_{\bar{U}}}} Q_{\bar{U}}(s, a, M) = B Q_{\bar{U}}(s, a, M) \), and in step 7 we used the fact that \( Q_{\bar{U}}(s, a, M) - B Q_{\bar{U}}(s, a, M) \leq \epsilon_a \).

Note that both the first half of lemma 13.5 (used in the first half of the proof) and the second half (used in the second half of the proof) hold simultaneously with probability \( \frac{11}{2} \), thus we don’t need to take a union bound over the individual probabilities.

\[ \square \]
Lemma 13.7 bounds the number of times we can encounter approximation units with fewer than \( k \) active samples.

**Lemma 13.7.** Let \((s_{1,j}, s_{2,j}, s_{3,j}, \ldots)\) for \( j \in \{1, \ldots, k_p\}\) be the random paths generated in MDPs \( M_1, \ldots, M_{k_p} \) respectively on some execution of Algorithm 1. Let \( \hat{U}(t,j) \) be the approximation set used by Algorithm 1 at step \( t \) in MDP \( M_j \). Let \( \tau(t,j) \) be the number of steps from step \( t \) in MDP \( M_j \) to the next delay step, or to the first step \( t' \) for which \( \hat{U}(t,j) \neq \hat{U}(t',j) \), whichever comes first.

Let \( T_H = \left\lfloor \frac{1}{\gamma} \ln \frac{Q_{\text{max}}}{\epsilon_s} \right\rfloor \) and define \( H = \{1, 2, 4, \ldots, 2^i\} \) where \( i \) is the largest integer such that \( 2^i \leq T_H \). Let \( k_a^- \) be the largest value in \( K_a \) that is strictly smaller than \( k_a \), or 0 if such a value does not exist. Let \( X_{k_a}(t,j) \) be the set of state-action-MDP triples at step \( t \) in MDP \( M_j \) such that no approximation unit with at least \( k_a \) active samples exists within \( d_{\text{known}} \) distance, and at least one approximation unit with at least \( k_a^- \) active samples exists within \( d_{\text{known}} \) distance \((k_a^- = 0 \text{ covers the case where an approximation unit with } 0 \text{ active samples exists within } d_{\text{known}} \text{ distance})\). Define \( p_{h,k_a}(s_{t,j}) \) for \( k_a \in K_a \) to be Bernoulli random variables that express the following conditional probability: Given the state of the approximation set and sample candidate queue at step \( t \) for MDP \( M_j \), exactly \( h \) state-action-MDP triples in \( X_{k_a}(t,j) \) are encountered in MDP \( M_j \) during the next \( \min\{T_H, \tau(t,j)\} \) steps. Let \( p_{h,k_a}(s_{t,j}) = \sum_{i=0}^{2^h-1} p_i(k_a)(s_{t,j}) \).

Finally let \( T_j \) be the set of non-delay steps (see Definition 7.5) for MDP \( M_j \).

If

\[
\frac{2^i \left\lfloor \frac{1}{1-\gamma} \ln \frac{Q_{\text{max}}}{\epsilon_s} \right\rfloor}{k_p N_\text{SAM}(d_{\text{known}})} \ln \frac{2^{\left\lfloor 1+\log_2 k \right\rfloor}}{\delta} < 1 \quad \text{and} \quad N_\text{SAM}(d_{\text{known}}) \geq 2,
\]

with probability \( 1 - \frac{\delta}{2^i} \):

\[
\sum_{j=1}^{k_p} \sum_{t \in T_j} \sum_{h \in H} \left( h p_{h,k_a}(s_{t,j}) \right) \leq \frac{(k_a - k_a^- + k_p) \left( 1 + \log_2 \left( \frac{1}{\gamma} \ln \frac{Q_{\text{max}}}{\epsilon_s} \right) \right) \left( 1 - \frac{1}{\gamma} \ln \frac{Q_{\text{max}}}{\epsilon_s} \right) N_\text{SAM}(d_{\text{known}})}{1 - \sqrt{\frac{2^i \left\lfloor \frac{1}{1-\gamma} \ln \frac{Q_{\text{max}}}{\epsilon_s} \right\rfloor}{k_p N_\text{SAM}(d_{\text{known}})} \ln \frac{2^{\left\lfloor 1+\log_2 k \right\rfloor}}{\delta} (k_a - k_a^- + k_p) N_\text{SAM}(d_{\text{known}})}}
\]

\( \forall k_a \in K_a \) and \( \forall h \in H \) simultaneously.

**Proof.** From the Markov property we have that, \( p_{h,k_a}(s_{t,j}) \) variables at least \( T_H \) steps apart (of the same or different \( j \)) are independent\(^3\). Define \( T_i^H \) for \( i \in \{0, 1, \ldots, T_H - 1\} \) to be the (finite) set of timesteps for which \( t \in \{i, i + T_H, i + 2T_H, \ldots\} \).

\( k_a - k_a^- \) samples will be added to an approximation unit with \( k_a^- \) active samples before it progresses to \( k_a \) active samples. Additionally, at most \( N_\text{SAM}(d_{\text{known}}) \) approximation units can be added to the approximation set, and we are exploring \( k_p \) MDPs in parallel. Thus, at most \((k_a - k_a^- + k_p - 1)N_\text{SAM}(d_{\text{known}})\)

\(^3\)Careful readers may notice that what happens at step \( t \) affects which variables are selected at future timesteps. This is not a problem. We only care that the outcomes of the variables are independent given their selection.
state-action-MDP triples within $d_{\text{known}}$ distance of approximation units with $k_a^-$ active samples can be encountered on non-delay steps\(^4\).

Let us assume that there exists an $i \in \{0, 1, \ldots, T_H - 1\}$ and $h \in H$ such that:

$$
\sum_{j=1}^{k_p} \sum_{t \in T_j \cap T_i^h} p_{h,k_a}(s_{t,j}) \geq \frac{(k_a - k_a^- + k_p)\mathcal{NSAM}(d_{\text{known}})}{h \left(1 - \sqrt{\frac{2h}{(k_a - k_a^- + k_p)\mathcal{NSAM}(d_{\text{known}})} \ln \frac{2(1 + \log_2 k^H)}{\delta}}\right)}.
$$

From lemma 13.1 it follows that with probability at least $1 - \frac{\delta}{2(1 + \log_2 k^H)}$, at least $(k_a - k_a^- + k_p)\mathcal{NSAM}(d_{\text{known}})$ state-action-MDP triples within $d_{\text{known}}$ distance of approximation units with $k_a^-$ active samples will be encountered on non-delay steps, which is a contradiction. It must therefore be the case that:

$$
\sum_{j=1}^{k_p} \sum_{t \in T_j \cap T_i^h} p_{h,k_a}(s_{t,j}) < \frac{(k_a - k_a^- + k_p)\mathcal{NSAM}(d_{\text{known}})}{h \left(1 - \sqrt{\frac{2h}{(k_a - k_a^- + k_p)\mathcal{NSAM}(d_{\text{known}})} \ln \frac{2(1 + \log_2 k^H)}{\delta}}\right)}
$$

with probability at least $1 - \frac{\delta}{2(1 + \log_2 k^H)}$ for all $i \in \{0, 1, \ldots, T_H - 1\}$ and $h \in H - \{T_H\}$ simultaneously, which implies that:

$$
\sum_{j=1}^{k_p} \sum_{t \in T_j} \sum_{h \in H} (h p_{h,k_a}(s_{t,j})) < 1 - \frac{\delta}{2(1 + \log_2 k^H)}
\sum_{j=1}^{k_p} \sum_{t \in T_j} \sum_{h \in H} (h p_{h,k_a}(s_{t,j}))
$$

with probability at least $1 - \frac{\delta}{2(1 + \log_2 k^H)}$ for all $h \in H$ simultaneously.

From the union bound we have that since $k_a$ can take $\lceil 1 + \log_2 k \rceil$ values, with probability $1 - \frac{\delta}{2}$:

$$
\sum_{j=1}^{k_p} \sum_{t \in T_j} \sum_{h \in H} (h p_{h,k_a}(s_{t,j})) < 1 - \frac{\delta}{2(1 + \log_2 k^H)}
\sum_{j=1}^{k_p} \sum_{t \in T_j} \sum_{h \in H} (h p_{h,k_a}(s_{t,j}))
$$

\[\forall \, k_a \in K_a \text{ and } \forall \, h \in H \text{ simultaneously.}\]

\[^4\text{This covers the worst case, which could happen if every time a state-action-MDP triple within } d_{\text{known}} \text{ distance of an approximation unit with } k_a - 1 \text{ samples is encountered by one of the MDPs, the remaining } k_p - 1 \text{ MDPs encounter a state-action-MDP triple within } d_{\text{known}} \text{ of the same approximation unit simultaneously.}\]
8 Proofs of main lemmas and theorems

Lemma 8.1. The space complexity of algorithm 1 is:

$$\mathcal{O}(N_{S,AM}(d_{\text{known}}))$$

per concurrent MDP.

Proof. Algorithm 1 only needs access to the value and number of active samples of each approximation unit, and from the definition of the covering number and Algorithm 2 we have that at most $N_{S,AM}(d_{\text{known}})$ approximation units can be added. \hfill \Box

Lemma 8.2. The space complexity of algorithm 2 is:

$$\mathcal{O}(k|A|N_{S,AM}(d_{\text{known}}))$$

Proof. From the definition of the covering number we have that at most $N_{S,AM}(d_{\text{known}})$ approximation units can be added. From the definition of an approximation unit and Algorithm 2 we have that at most $k$ samples can be added per approximation unit, and each sample requires at most $\mathcal{O}(|A|)$ space. \hfill \Box

Lemma 8.3. The per step computational complexity of algorithm 1 is bounded above by:

$$\mathcal{O}(|A|N_{S,AM}(d_{\text{known}})).$$

Proof. A naive search for the nearest approximation unit of each of the at most $|A|$ actions needs to perform at most $\mathcal{O}(N_{S,AM}(d_{\text{known}}))$ operations. \hfill \Box

Since every step of algorithm 1 results in a sample being processed by algorithm 2, we will be bounding the per sample computational complexity of algorithm 2:

Lemma 8.4. Let $c$ be the maximum number of approximation units to which a single sample can be added\(^5\). The per sample computational complexity of algorithm 2 is bounded above by:

$$\mathcal{O}\left(ck|A|N_{S,AM}(d_{\text{known}}) + \frac{k|A|N_{S,AM}(d_{\text{known}})}{\ln \gamma} \ln \frac{\epsilon_a}{Q_{\text{max}}}\right).$$

Proof. Lines 4 through 12 will be performed only once per sample. Line 6 requires at most $\mathcal{O}(N_{S,AM}(d_{\text{known}}))$ operations. Line 7 requires at most $\mathcal{O}(kN_{S,AM}(d_{\text{known}}))$ operations. Line 9 can reuse the result of line 6 and only check if the approximation unit added in line 7 (if any) has $k$ samples. Line 10 requires at most $\mathcal{O}(N_{S,AM}(d_{\text{known}}))$ operations.

\(^5c\) will depend on the dimensionality of the state-action-MDP space. For domains that are big enough to be interesting it will be significantly smaller than other quantities of interest.
Given a pointer to $N(U, s', a', M)$ for every next state action MDP triple, line 14 requires at most $O(k|A|N_{SA,M}(d_{known}))$ operations ($O(k|A|)$ for each approximation unit), and since $B$ is a $\gamma$-contraction in maximum norm (lemma 13.2), it will be performed at most $\frac{\ln \frac{Q_{max}}{\varepsilon}}{\ln \gamma}$ times per sample\(^6\) (lemma 13.3), for a total cost of $O\left(\frac{k|A|N_{SA,M}(d_{known})}{\ln \gamma} \ln \frac{\varepsilon}{Q_{max}}\right)$. This step dominates the computational cost of the learner.

Lines 15 and 16 require at most $O(N_{SA,M}(d_{known}))$ operations each, and will be performed at most $\frac{\ln \frac{Q_{max}}{\varepsilon}}{\ln \gamma}$ times per sample.

Maintaining a pointer to $N(U, s', a', M)$ for every next state action MDP triple requires the following: 1) Computing $N(U, s', a', M)$ every time a sample is added to the sample set, for a cost of up to $O(|A|N_{SA,M}(d_{known}))$ operations per sample. 2) Updating all $N(U, s', a', M)$ pointers every time the number of active samples of an approximation unit changes: There can exist at most $k|A|N_{SA,M}(d_{known})$ pointers in the sample set, and a single sample can affect the number of active samples of at most $c$ approximation units. Each pointer can be updated against a particular approximation unit in constant time, thus the cost of maintaining the pointers is at most $O(ck|A|N_{SA,M}(d_{known}))$ operations per sample.

\(\square\)

**Theorem 8.5.** Let \(s_{1,j}, s_{2,j}, s_{3,j}, \ldots\) for \(j \in \{1, \ldots, k_p\}\) be the random paths generated in MDPs \(M_1, \ldots, M_{k_p}\) respectively on some execution of algorithm 1, and \(\tilde{\pi}\) be the (non-stationary) policy followed by algorithm 1. Let 
\[
\varepsilon_b = \hat{Q}_{max} \sqrt{\frac{\ln \frac{4[1 + \log_2 k|A|N_{SA,M}(d_{known})]^2}{2}}{4[1 + \log_2 k|A|N_{SA,M}(d_{known})]^2}} \cdot k \geq \frac{Q_{max}^2}{2 \gamma^2 (1 - \gamma)^2} \ln \left(\frac{4[1 + \log_2 k|A|N_{SA,M}(d_{known})]^2}{\delta}\right),
\]
\(\varepsilon_c\) be defined as in definition 5.13, \(\varepsilon_a\) be defined as in algorithm 2, \(k_p\) be defined as in definition 7.1, and \(T_j\) be the set of non-delay steps (see definition 7.5) for MDP \(M_j\). If 
\[
\frac{2[1 + \log_2 k|A|N_{SA,M}(d_{known})]}{(1 + k_p)[N_{SA,M}(d_{known})]} < 1 \quad \text{and} \quad N_{SA,M}(d_{known}) \geq 2,
\]
with probability at least \(1 - \delta\), for all \(t, j\):
\[
V^\pi(s_{t,j}, M_j) \geq V^\ast(s_{t,j}, M_j) - \frac{4\varepsilon_c + 2\varepsilon_a}{1 - \gamma} - 3\varepsilon_a - \varepsilon_c(t, j),
\]
where\(^7\):
\[
\sum_{j=1}^{k_p} \sum_{t=0}^{\infty} \varepsilon_c(t, j) = \sum_{j=1}^{k_p} \sum_{t \in T_j} \varepsilon_c(t, j) + \sum_{j=1}^{k_p} \sum_{t \notin T_j} \varepsilon_c(t, j),
\]
with:
\[
\sum_{j=1}^{k_p} \sum_{t \in T_j} \varepsilon_c(t, j)
\]

\(^6\)Note that the per sample cost does not increase even if multiple samples enter through line 4 before publishing $\hat{U}$.

\(^7\)\(f(n) = O(g(n))\) is a shorthand for \(f(n) = O(g(n) \log^c g(n))\) for some $c$. 

15
\[ \approx \hat{O} \left( \frac{\hat{Q}_{\text{max}}}{\epsilon_{t(1-\gamma)}} + k_p N_S A_M (d_{\text{known}}) \hat{Q}_{\text{max}} \right) \]

and:

\[ \sum_{t \in T_j} \epsilon_e(t, j) \leq \left( Q_{\text{max}} \sum_{k_a \in K_a} \sum_{t=1}^{N_S A_M (d_{\text{known}})} D_{i,j,k_a} \right), \]

where \( D_{i,j,k_a} \) is the delay of the \( k_a \)-th sample for \( k_a \in K_a \) in the \( i \)-th approximation unit, with respect to MDP \( M_j \). Note that the probability of success \( 1 - \delta \) holds for all timesteps in all MDPs simultaneously, and \( \sum_{j=1}^{k_p} \sum_{t=0}^{\infty} \epsilon_e(t, j) \) is an undiscounted infinite sum. Unlike \( \sum_{j=1}^{k_p} \sum_{t \in T_j} \epsilon_e(t, j) \) which is a sum over all MDPs, \( \sum_{t \in T_j} \epsilon_e(t, j) \) gives a bound on the total cost due to value function update delays for each MDP separately.

Proof. We have that:

\[ Q_G(s,a,M) - B^{\pi,\tilde{Q}} Q_G(s,a,M) \leq Q_{\text{max}} \forall (s,a,M,\tilde{U}). \]

Since \( k_a^- \geq k_a^- \forall k_a \in K_a - \{1\} \) we have that for any \( (s,a,M) \) within \( d_{\text{known}} \) distance of an approximation unit with \( k_a^- > 0 \) active samples, with probability \( 1 - \frac{\delta}{2} \):

\[ Q_G(s,a,M) - B^{\pi,\tilde{Q}} Q_G(s,a,M) < \epsilon_a + 2\epsilon_e + 2\hat{Q}_{\text{max}} \left( \frac{\ln \left( \frac{d_{\text{known}}}{\delta} \right)}{k_a} \right). \]

Substituting \( k \geq \frac{\hat{Q}_{\text{max}}^2}{2\epsilon_{t(1-\gamma)}} \ln \left( \frac{d_{\text{known}}}{\delta} \right) \) for \( k_a \) into lemma 13.6, we have that with probability at least \( 1 - \frac{\delta}{2} \) for any \( (s,a,M) \) within \( d_{\text{known}} \) of an approximation unit with \( k \) active samples\(^8\):

\[ \hat{Q}(s,a,M) - B^{\pi,\tilde{Q}} \hat{Q}(s,a,M) < \epsilon_a + 2\epsilon_e + 2(1-\gamma)\epsilon_s. \]

Let \( T_H, H, \tilde{U}(t,j), \tau(t,j), \) and \( p_{\tilde{h},k_a}^e(s_{t,j}) \) be defined as in lemma 13.7. Even though \( \tilde{\pi} \) is non-stationary, it is comprised of stationary segments. Starting from step \( t \) in MDP \( M_j \), \( \tilde{\pi} \) is equal to \( \pi^{Q_{\text{max}},t(j)} \) for at least \( \tau(t,j) \) steps. Substituting the above into lemma 12.6 we have that with probability at least \( 1 - \frac{\delta}{2} \), for all \( M_j \) and \( t \in T_J \):

\[ V^*(s_{t,j},M_j) - V^{\tilde{\pi}}(s_{t,j},M_j) \leq \frac{4\epsilon_e + 2\epsilon_a}{1-\gamma} + 3\epsilon_s + \epsilon_e(t,j), \]

where:

\[ \epsilon_e(t,j) = \gamma^{\tau(t,j)} Q_{\text{max}} + 2 \sum_{h \in H} (hp_{h,1}^e(s_{t,j}))Q_{\text{max}} \]

\(8\) Notice that in this case \( d(s, a, M, N(s, a, M), d_{\text{known}}) = 0 \).
Define:

\[ c_0 = \left(1 + \log_2 \left[ \frac{1 - \gamma}{1 - \gamma} \ln \frac{Q_{\text{max}}}{\epsilon_s} \right] \right) \left[ \frac{1}{1 - \gamma} \ln \frac{Q_{\text{max}}}{\epsilon_s} \right] N_{SAM}(d_{\text{unknown}}). \]

From the above it follows that:

\[ \sum_{j=1}^{k_p} \sum_{t \in T_j} \alpha(t,j) \]

\[ = \sum_{j=1}^{k_p} \sum_{t \in T_j} \gamma(t,j) Q_{\text{max}} \]

\[ + 2 \sum_{h \in H} (h \rho_{h,1}(s,t,j)) Q_{\text{max}} \]

\[ + \sum_{k_a \in (K_a - 1)} \sum_{j=1}^{k_p} \sum_{t \in T_j} (h \rho_{h,k_a}(s,t,j)) \sqrt{\frac{2 \ln \frac{4[1+\log_2 k]N_{SAM}(d_{\text{unknown}})^2}{k_a}}{k_a}} \]

\[ \leq k_p \left[ 1 + \log_2 k \right] N_{SAM}(d_{\text{unknown}}) Q_{\text{max}} \]

\[ \leq k_p \left[ 1 + \log_2 k \right] N_{SAM}(d_{\text{unknown}}) Q_{\text{max}} \]

\[ + \sum_{k_a \in (K_a - 1)} \sum_{j=1}^{k_p} \sum_{t \in T_j} (h \rho_{h,k_a}(s,t,j)) \sqrt{\frac{2 \ln \frac{4[1+\log_2 k]N_{SAM}(d_{\text{unknown}})^2}{k_a}}{k_a}} \]

\[ \leq k_p \left[ 1 + \log_2 k \right] N_{SAM}(d_{\text{unknown}}) Q_{\text{max}} \]

\[ + \sum_{k_a \in (K_a - 1)} \sum_{j=1}^{k_p} \sum_{t \in T_j} (h \rho_{h,k_a}(s,t,j)) \sqrt{\frac{2 \ln \frac{4[1+\log_2 k]N_{SAM}(d_{\text{unknown}})^2}{k_a}}{k_a}} \]

\[ \leq k_p \left[ 1 + \log_2 k \right] N_{SAM}(d_{\text{unknown}}) Q_{\text{max}} \]
\[
\begin{align*}
&\leq \frac{(2 + [3 + \log_2 k]\kappa_p)C_0Q_{\text{max}}}{1 - \sqrt{\frac{1}{2} \frac{1}{1 - \gamma} \ln \frac{Q_{\text{max}}}{\epsilon_s} \ln \frac{2[1 + \log_2 k]N_{\text{SAAM}}(d_{\text{known}})}{\delta}} + 2\sum_{k_d \in (K_{\text{a}} - 1)} \left( \sqrt{k_d} + \frac{2k_d}{\sqrt{2k_d}} \right) C_0Q_{\text{max}} \sqrt{\frac{2\ln \frac{4[1 + \log_2 k]N_{\text{SAAM}}(d_{\text{known}})^2}{\delta}}{1 - \sqrt{\frac{1}{2} \frac{1}{1 - \gamma} \ln \frac{Q_{\text{max}}}{\epsilon_s} \ln \frac{2[1 + \log_2 k]N_{\text{SAAM}}(d_{\text{known}})}{\delta}}}}
\end{align*}
\]

with probability \(1 - \delta\), where in step 3 we used the fact that there can be at most \([1 + \log_2 k]N_{\text{SAAM}}(d_{\text{known}})\) policy changes and lemma 13.7. And in step 8 we used the fact that \(Q_{\text{max}}\sqrt{2\ln \frac{4[1 + \log_2 k]N_{\text{SAAM}}(d_{\text{known}})^2}{\delta}} \geq Q_{\text{max}}\). Since lemma 13.6 (used to bound the Bellman error of each \((s, a, M, U)\)) and lemma 13.7 (used to bound how many times each \((s, a, M, U)\) is encountered) hold with probability at least \(1 - \frac{\delta}{2}\) each, the bound above holds with probability at least \(1 - \delta\).

Since for realistically large \(N_{\text{SAAM}}(d_{\text{known}})\):

\[
1 - \sqrt{\frac{2\frac{1}{1 - \gamma} \ln \frac{Q_{\text{max}}}{\epsilon_s} \ln \frac{2[1 + \log_2 k]}{\delta}}{1 + k_pN_{\text{SAAM}}(d_{\text{known}})}} \approx 1,
\]
we have that:

\[
\sum_{j=1}^{k_p} \sum_{t \in T_j} \epsilon_c(t, j) < \left( \left( 1 + \log_2 \left[ \frac{1}{1 - \gamma} \frac{Q_{\text{max}}}{\epsilon_s} \right] \right) \left[ \frac{1}{1 - \gamma} \frac{Q_{\text{max}}}{\epsilon_s} \right] N_{\text{SAM}}(d_{\text{known}}) \hat{Q}_{\text{max}} \right) \frac{2 + [3 + \log_2 k]k_p}{\epsilon e p t, j Q_{\text{max}}} \left( 9 \sqrt{F} + 10k_p \right) \sqrt{2 \ln \left( 2 \left[ 1 + \log_2 k \right] N_{\text{SAM}}(d_{\text{known}})^2 \right)}
\]

\[
= O \left( \frac{Q_{\text{max}}}{\epsilon(1 - \gamma)} + k_p N_{\text{SAM}}(d_{\text{known}}) \hat{Q}_{\text{max}} \right)
\]

with probability \(1 - \delta\), where we have substituted \(k\) with \(c\hat{Q}_{\text{max}}^2 / 2\epsilon^2(1 - \gamma)^2 \ln \left( 4[1 + \log_2 k] N_{\text{SAM}}(d_{\text{known}})^2 \right) \delta^{-1}\) for some constant \(c\).

From the definition of a delay step (Definition 7.5) we have that the maximum number of delay steps in MDP \(M_j\) is \(\sum_{k_a \in K_a} \sum_{i=1}^{N_{\text{SAM}}(d_{\text{known}})} D_{i,j,k_a}\). Allowing for the algorithm to perform arbitrarily badly (\(\epsilon_c(t, j) = Q_{\text{max}}\)) on every delay step we have that \(\forall j:\)

\[
\sum_{t \notin T_j} \epsilon_c(t, j) \leq Q_{\text{max}} \sum_{k_a \in K_a} \sum_{i=1}^{N_{\text{SAM}}(d_{\text{known}})} D_{i,j,k_a}.
\]

\[\square\]

References
