Supplementary Material

Submission 7268

1 Packing and Covering Numbers

For completeness, we briefly review the concepts of δ-packing and -covering numbers in general metric spaces. Let $B(x, \delta)$ denote the closed ball of radius δ centred at x.

Definition 1. *Let (X, ρ) be a metric space, let Y be a subset of X. For $\delta > 0$, the set of points $\{x_1, \dots, x_n\} \subset X$ is a* δ-covering *of Y if $Y \subset \bigcup_{i=1}^n B(x_i, \delta)$, or equivalently, $\forall y \in Y$, $\exists i$ such that $\rho(y, x_i) \leq \delta$. If, moreover, the set $\{x_1, \dots x_n\}$ is a subset of Y, then we say it is an* internal δ-covering *of Y.*[1] *Respectively, the* δ-covering number *and* internal δ-covering number *are defined as*

$$\mathcal{C}_\delta(Y) := \inf\{n : \exists \text{ a } \delta\text{-covering of } Y \text{ of size } n\} \quad (1)$$

and

$$\mathcal{C}^\circ_\delta(Y) := \inf\{n : \exists \text{ an internal } \delta\text{-covering of } Y \text{ of size } n\}. \quad (2)$$

Definition 2. *Let (X, ρ) be a metric space. For $\delta > 0$ and $Y \subset X$, the set of points $\{y_1, \dots, y_m\} \subset Y$ is a* δ-packing *of Y if, $\forall i \neq j$, $\rho(x_i, x_j) > \delta$ (notice strict inequality), or equivalently, $\bigcap_{i=1}^m B(y_i, \delta) = \emptyset$. The* δ-packing number *is defined as*

$$\mathcal{P}_\delta(Y) := \sup\{m : \exists \text{a } \delta\text{-packing of size } m\}. \quad (3)$$

Proposition 1 (Packing-Covering Duality)**.** *Let (X, ρ) be a metric space, and $Y \subset X$. Then for arbitrary $\delta > 0$*

$$\mathcal{P}_\delta(Y) \leq \mathcal{C}_{\delta/2}(Y) \leq \mathcal{C}^\circ_{\delta/2}(Y) \leq \mathcal{P}_{\delta/2}(Y). \quad (4)$$

Proof. To prove the first inequality suppose, for a contradiction, that there exists a δ-packing $\{y_1, \dots, y_m\}$ and a $\delta/2$-covering $\{x_1, \dots, x_n\}$ such that $m > n$. By the pigeonhole principle, there must be at least one pair y_i and y_j belonging to the same ball $B(x_k, \delta/2)$ for some k. This means that $\rho(y_i, y_j) \leq \delta$ thereby contradicting the fact that $\{y_1, \dots, y_m\}$ is a δ-packing. Hence $m \leq n$ and the conclusion follows.

For the last inequality, let $\mathcal{E} = \{y_1, \dots, y_m\} \subset Y$ be a maximal packing. Suppose, for a contradiction, that there is a $y \in Y \setminus \mathcal{E}$ such that $\forall i$ we have $\rho(y_i, y) > \delta/2$. But this contradicts the maximality of $\mathcal{E}$ since we can simply construct a larger packing with $y_{m+1} := y$. Hence, $\mathcal{E}$ must be an internal $\delta/2$-covering of Y. Since $\mathcal{C}^\circ_{\delta/2}(Y)$ is the minimal size of all possible $\delta/2$-coverings, we have $\mathcal{C}^\circ_{\delta/2}(Y) \leq \mathcal{P}_{\delta/2}(Y)$.

The middle inequality follows trivially by observing that the set of internal δ-coverings is a subset of the set of all δ-coverings. ☐

Remark 1. *The quantity $\mathcal{C}_\delta(Y)$ is finite for all $\delta > 0$ if and only if Y is* totally bounded *– i.e. for every $\delta > 0$, there is a finite covering of the space by balls of radius δ. It is a well known fact that a compact set is totally bounded. Moreover, a bounded subset of a Euclidean space of arbitrary (finite) dimension is totally bounded (see e.g. [Kolmogorov and Fomin, 1970]). Therefore, if $|\mathcal{S}| < \infty$, the set $\mathcal{R}_{b_0}$ is totally bounded, being bounded by the $(|\mathcal{S}| - 1)$-dimensional simplex in $\mathbb{R}^{|\mathcal{S}|}$. This observation, along with Proposition 1, means that the quantities $\mathcal{C}_\delta(\mathcal{R}_{b_0})$, $\mathcal{C}^\circ_\delta(\mathcal{R}_{b_0})$ and $\mathcal{P}_\delta(\mathcal{R}_{b_0})$ are all well defined (i.e. finite) in this case. However, when $|\mathcal{S}| = \infty$ the finiteness of $\mathcal{C}_\delta(\mathcal{R}_{b_0})$ is not guaranteed in general. We therefore assert that $\mathcal{R}_{b_0}$ is totally bounded as per Assumption 1 in the main body of the paper.*

2 Some Useful Operators and Results

We will find it convenient to introduce some function operators for notational compactness. Recall that the operator $\tilde{\tau}_{B,\rho}$ was introduced in the main body of the paper.

Definition 3. *Let $\mathscr{V}$ and $\mathscr{Q}$ respectively denote the space of all functions $v : \mathcal{B} \to \mathbb{R}$ and $q : \mathcal{B} \times \mathcal{A} \to \mathbb{R}$ such that $\|v\|_\infty < \infty$ and $\|q\|_\infty < \infty$. For a given subset $B \subset \mathcal{B}$ and (stochastic) policy $\pi : \mathcal{B} \to \Delta(\mathcal{A})$ let*

$$\langle \pi, q \rangle(b) := \sum_{a \in \mathcal{A}} \pi(a \,|\, b) q(b, a) \quad (5)$$

$$[\mathcal{M}q](b) := \max_{a \in \mathcal{A}} q(b, a) \quad (6)$$

$$\langle P, v \rangle(b, a) := \sum_{o \in \mathcal{O}} P(o \,|\, b, a) v\big(\tau(b, a, o)\big) \quad (7)$$

$$\langle P, v \rangle_{B,\rho}(b, a) := \sum_{o \in \mathcal{O}} P(o \,|\, b, a) v\big(\tilde{\tau}_{B,\rho}(b, a, o)\big) \quad (8)$$

where $\tilde{\tau}_{B,\rho}$ was defined in Sect. 4.3 of the paper.

The next result is a simple consequence of Definition 3.

[1] In [Lee *et al.*, 2007], such a covering is called a *proper δ covering*. Our terminology seems more descriptive and is consistent with the broader mathematical literature.

Proposition 2. *For any $B \subset \mathcal{B}$, metric ρ on $\mathcal{B}$, $v_1, v_2 \in \mathscr{V}$ and $q_1, q_2 \in \mathscr{Q}$ and $\mu_1, \mu_2, \lambda_1, \lambda_2 \in \mathbb{R}$ we have*

$$\begin{aligned}&\langle \pi, \mu_1 v_1 + \mu_2 v_2 + \lambda_1 q_1 + \lambda_2 q_2 \rangle \\ &\qquad = \mu_1 v_1 + \mu_2 v_2 + \lambda_1 \langle \pi, q_1 \rangle + \lambda_2 \langle \pi, q_2 \rangle \end{aligned} \tag{9}$$

and

$$\begin{aligned}&\langle P, \langle \pi, \mu_1 v_1 + \mu_2 v_2 + \lambda_1 q_1 + \lambda_2 q_2 \rangle \rangle_{B,\rho} \\ &\qquad = \mu_1 \langle P, v_1 \rangle_{B,\rho} + \mu_2 \langle P, v_2 \rangle_{B,\rho} \\ &\qquad + \lambda_1 \langle P, \langle \pi, q_1 \rangle \rangle_{B,\rho} + \lambda_2 \langle P, \langle \pi, q_2 \rangle \rangle_{B,\rho}. \end{aligned} \tag{10}$$

Moreover

$$\| \langle P, v_1 \rangle_{B,\rho} \|_\infty \leq \| v_1 \|_{B,\infty}. \tag{11}$$

Proof. Equations (9) and (10) follow straightforwardly from the definitions. The inequality (11) follows from Jensen's inequality since

$$\begin{aligned}|\langle P, v_1 \rangle_{B,\rho}(b,a)| &= \Big| \sum_{o \in \mathcal{O}} P(o \,|\, b, a) v_1\big(\tilde{\tau}_{B,\rho}(b,a,o)\big) \Big| \\ &\leq \sum_{o \in \mathcal{O}} P(o \,|\, b, a) \big| v_1\big(\tilde{\tau}_{B,\rho}(b,a,o)\big) \big| \leq \| v_1 \|_{B,\infty}\end{aligned}$$

for every $b \in \mathcal{B}$ and $a \in \mathcal{A}$. ☐

For a given stochastic policy $\pi \in \Pi$, let the self-mapping operators $\mathscr{T} : \mathscr{Q} \to \mathscr{Q}$, $\mathscr{T}^\pi : \mathscr{Q} \to \mathscr{Q}$ and $\mathscr{T}^\pi_{B,\rho} : \mathscr{Q} \to \mathscr{Q}$ be given by

$$[\mathscr{T} q](b,a) := R(b,a) + \gamma \langle P, [\mathcal{M} q] \rangle (b,a) \tag{12}$$

$$[\mathscr{T}^\pi q](b,a) := R(b,a) + \gamma \langle P, \langle \pi, q \rangle \rangle (b,a) \tag{13}$$

$$[\mathscr{T}^\pi_{B,\rho} q](b,a) := R(b,a) + \gamma \langle P, \langle \pi, q \rangle \rangle_{B,\rho} (b,a). \tag{14}$$

We have the following well-known result, which can be justified using a classical argument – see e.g. [Bertsekas, 2008] or [Ross, 1970].

Lemma 1. *For arbitrary q and q' belonging to $\mathscr{Q}$, the operators $\mathscr{T}$, $\mathscr{T}^\pi$ and $\mathscr{T}^\pi_{B,\rho}$ satisfy*

$$\begin{aligned}\| \mathscr{T} q - \mathscr{T} q' \|_\infty &\leq \gamma \| q - q' \|_\infty \\ \| \mathscr{T}^\pi q - \mathscr{T}^\pi q' \|_\infty &\leq \gamma \| q - q' \|_\infty \\ \| \mathscr{T}^\pi_{B,\rho} q - \mathscr{T}^\pi_{B,\rho} q' \|_\infty &\leq \gamma \| q - q' \|_\infty. \end{aligned} \tag{15}$$

Thus, for $\gamma \in (0,1)$, the operators are contraction mappings with modulus of contraction γ and have unique fixed points (up to $\| \cdot \|_\infty$-equivalence). In particular, $Q^*, Q^\pi, Q^\pi_{B,\rho} \in \mathscr{Q}$ are the fixed points of $\mathscr{T}$, $\mathscr{T}^\pi$ and $\mathscr{T}^\pi_{B,\rho}$ respectively. Moreover, Q^* is the solution of the POMDP introduced in Sec. 2.1.

We can quantify the difference between the fixed points when B is a δ-covering $\mathcal{E}_\delta$ of $\mathcal{R}_{b_0}$ for the metric ρ_1. We will need two auxiliary results to prove this claim which we formalise in Proposition 3. The first is essentially Lemma 1 from [Lee *et al.*, 2007].

Lemma 2. *For any $\delta > 0$ and belief points b and b', we have*

$$|V^*(b) - V^*(b')| \leq Q_{\max} \| b - b' \|_1 \tag{16}$$

and, for fixed $a \in \mathcal{A}$,

$$|Q^*(b,a) - Q^*(b',a)| \leq Q_{\max} \| b - b' \|_1. \tag{17}$$

Proof. The inequality (16) comes directly from [Lee *et al.*, 2007] We sketch the proof for (17) since it follows in a similar way to Lemma 1 in [Lee *et al.*, 2007]. For a fixed $a \in \mathcal{A}$, the function $Q^* : \mathcal{B} \times \mathcal{A} \to \mathbb{R}$ can be approximated arbitrarily closely by a piecewise-linear function $Q^*(b,a) = \max_{\alpha \in \Gamma} (\alpha \cdot b)$ where $\Gamma \subset \mathbb{R}^{|\mathcal{S}|}$. For each $\alpha \in \Gamma$ the boundedness of the reward function ensures that the absolute values of the components of α are bounded by $Q_{\max}$. We can then argue in a similar way to [Lee *et al.*, 2007] to get the desired bound. ☐

Lemma 3. *Consider any stochastic policy $\pi \in \Pi$ and δ-covering $\mathcal{E}_\delta$ of $\mathcal{R}_{b_0}$ for some $\delta > 0$. Then*

$$\| \mathscr{T}^\pi Q^* - \mathscr{T}^\pi_{\mathcal{E}_\delta, \rho_1} Q^* \|_\infty \leq \gamma \delta Q_{\max}. \tag{18}$$

Proof. For any $\pi \in \Pi$, and $q \in \mathscr{Q}$ and fixed $(b,a) \in \mathcal{B} \times \mathcal{A}$, the operator definitions and Lemma 2 give us

$$\begin{aligned}&\big| \mathscr{T}^\pi Q^*(b,a) - \mathscr{T}^\pi_{\mathcal{E}_\delta,\rho_1} Q^*(b,a) \big| \\ &\leq \gamma \Big| \big\langle P, \langle \pi, Q^* \rangle \big\rangle (b,a) - \big\langle P, \langle \pi, Q^* \rangle \big\rangle_{\mathcal{E}_\delta,\rho_1} (b,a) \Big| \\ &\leq \gamma \sum_{o,a'} \theta^{o,a'}_{b,a} \big| Q^*(\tau(b,a,o),a') - Q^*(\tilde{\tau}_{\mathcal{E}_\delta,\rho_1}(b,a,o),a') \big| \\ &\leq \gamma \sum_{o,a'} \theta^{o,a'}_{b,a} Q_{\max} \big\| \tau(b,a,o) - \tilde{\tau}_{\mathcal{E}_\delta,\rho_1}(b,a,o) \ \big\|_1 \\ &\leq \gamma \delta Q_{\max}\end{aligned}$$

where $\theta^{o,a'}_{b,a} := P(o \,|\, b,a) \pi(a' \,|\, b)$ is a probability distribution over $\mathcal{O} \times \mathcal{A}$. The desired inequality follows since (b,a) was arbitrary in $\mathcal{B} \times \mathcal{A}$. ☐

Proposition 3. *For any $\delta > 0$*

$$\| Q^* - Q^{\pi^*}_{\mathcal{E}_\delta,\rho_1} \|_\infty \leq \frac{\gamma \delta Q_{\max}}{1 - \gamma}. \tag{19}$$

Proof. The policy was arbitrary in Lemma 3, so we can choose $\pi := \pi^*$ and we get

$$\begin{aligned}&\| Q^* - Q^{\pi^*}_{\mathcal{E}_\delta,\rho_1} \|_\infty \\ &\leq \| \mathscr{T}^{\pi^*} Q^* - \mathscr{T}^{\pi^*}_{\mathcal{E}_\delta,\rho_1} Q^* + \mathscr{T}^{\pi^*}_{\mathcal{E}_\delta,\rho_1} Q^* - \mathscr{T}^{\pi^*}_{\mathcal{E}_\delta,\rho_1} Q^{\pi^*}_{\mathcal{E}_\delta,\rho_1} \|_\infty \\ &\leq \gamma \delta Q_{\max} + \gamma \| Q^* - Q^{\pi^*}_{\mathcal{E}_\delta,\rho_1} \|_\infty\end{aligned}$$

where we have used the contraction property from Lemma 1. Rearranging the above gives us the desired result. ☐

Definition 4. *For any $\eta > 0$ and $q \in \mathscr{Q}$, define the operators $\mathcal{L}_\eta : \mathscr{Q} \to \mathscr{V}$ and $\mathcal{M}_\eta : \mathscr{Q} \to \mathscr{V}$ according to*

$$[\mathcal{L}_\eta q](b) := \frac{1}{\eta} \log \Big\{ \sum_{a \in \mathcal{A}} \exp[\eta q(b,a)] \Big\} \tag{20}$$

$$[\mathcal{M}_\eta q](b) := \frac{\sum_{a \in \mathcal{A}} \exp[\eta q(b,a)] q(b,a)}{\sum_{a' \in \mathcal{A}} \exp[\eta q(b,a')]} \tag{21}$$

for all $b \in \mathcal{B}$ so that $\mathcal{L}_\eta$ and $\mathcal{M}_\eta$ are the log-sum-exp *and* Boltzmann soft-max *operators respectively.*

We have the following useful properties.

Proposition 4. *For any $v \in \mathscr{V}$ and $q \in \mathscr{Q}$ we have*

$$[\mathcal{L}_\eta(v+q)] = v + [\mathcal{L}_\eta q]. \tag{22}$$

Moreover, if $\mathcal{A}$ is finite,

$$\Big|[\mathcal{L}_\eta q](b) - [\mathcal{M}_\eta q](b)\Big| \le \frac{\log(|\mathcal{A}|)}{\eta} \tag{23}$$

and

$$0 \le [\mathcal{L}_\eta q](b) - [\mathcal{M}q](b) \le \frac{\log(|\mathcal{A}|)}{\eta} \tag{24}$$

for all $b \in \mathcal{B}$ and $\eta > 0$.

Proof. See [MacKay, 2003] for the proof of (23). We now prove (24). Fix a $b \in \mathcal{B}$. To show the lower bound, suppose $a^* \in \arg\max_{a\in\mathcal{A}} q(b,a)$ so that $q(b,a^*) = [\mathcal{M}q](b)$. Then

$$\begin{aligned}[\mathcal{L}_\eta q](b) &= \frac{1}{\eta}\log\Big\{\sum_{a\in\mathcal{A}}\exp[\eta q(b,a)]\Big\}\\ &\ge \frac{1}{\eta}\log\Big\{\exp[\eta q(b,a^*)]\Big\} = [\mathcal{M}q](b).\end{aligned}$$

Observe also that

$$\begin{aligned}[\mathcal{L}_\eta q](b) &\le \frac{1}{\eta}\log\Big\{\sum_{a\in\mathcal{A}}\exp\big[\eta[\mathcal{M}q](b)\big]\Big\}\\ &\le \frac{\log(|\mathcal{A}|)}{\eta} + [\mathcal{M}q](b)\end{aligned}$$

which is the desired upper bound. □

The following basic result will be useful.

Proposition 5. *Let $x \in X$ be an arbitrary element of a general set X and consider the real functions $f_1 : X \to \mathbb{R}$ and $f_2 : X \to \mathbb{R}$. Then*

$$\begin{aligned}&\sup_{x\in X} f_1(x) + \sup_{x\in X} f_2(x)\\ &\quad\le \sup_{x\in X}\Big(f_1(x)+f_2(x)\Big) + 2\Big(\sup_{x\in X}|f_2(x)|\Big).\end{aligned} \tag{25}$$

Proof. Observe that

$$\begin{aligned}f_1(x_1)+f_2(x_2) &= f_1(x_1)+f_2(x_1)-f_2(x_1)+f_2(x_2)\\ &= \sup_{x\in X}\Big(f_1(x)+f_2(x)\Big) + |f_2(x_1)| + |f_2(x_2)|\\ &= \sup_{x\in X}\Big(f_1(x)+f_2(x)\Big) + 2\Big(\sup_{x\in X}|f_2(x)|\Big)\end{aligned}$$

for arbitrary $x_1, x_2 \in X$. □

3 Convergence of the Exact Scheme

We can now prove Theorem 1 in the main body of the paper. In fact the proof is slightly more general as it states the result for all $B \subset \mathcal{R}_{b_0}$ (NB: Theorem 1 is the special case when $B = \mathcal{R}_{b_0}$). To this end, let $\{Q_0, Q_1, Q_2, \ldots\} \subset \mathscr{Q}$ be an auxiliary sequence of action-value functions which is defined according to recursion

$$Q_0 := \hat{\Psi}_0 \tag{26}$$

$$Q_k := R + \frac{\gamma}{k}\langle P, \mathcal{L}_\eta[(k-1)Q_{k-1}+Q_0]\rangle_{B,\rho} + \frac{E_{k-1}}{k} \tag{27}$$

for $k \ge 1$. Our aim is to bound the quantity $\|Q^{\hat{\pi}_k}_{B,\rho} - Q_k\|_\infty$ which will help us ultimately bound $\|Q^{\pi^*}_{B,\rho} - Q^{\hat{\pi}_k}_{B,\rho}\|_\infty$.

Step 1. In the first step, we will inductively verify the relation:

$$\hat{\Psi}_k = kQ_k + Q_0 - \mathcal{L}_\eta[(k-1)Q_{k-1}+Q_0], \quad \forall k \ge 1. \tag{28}$$

For the base case when $k = 1$, notice that (27) gives $Q_1 = R + \gamma\langle P, \mathcal{L}_\eta Q_0\rangle_{B,\rho} + \epsilon_0$. Then, the RHS of (28) is

$$\begin{aligned}&R + \gamma\langle P, \mathcal{L}_\eta Q_0\rangle_{B,\rho} + Q_0 - \mathcal{L}_\eta Q_0 + \epsilon_0\\ &\quad= R + \gamma\langle P, \mathcal{L}_\eta Q_0\rangle_{B,\rho} + \hat{\Psi}_0 - \mathcal{L}_\eta Q_0 + \epsilon_0 = \hat{\Psi}_1\end{aligned} \tag{29}$$

because of the synchronous scheme. For the induction step, suppose (28) holds up to some $k \ge 1$. Then, using Proposition 2 and our definitions, we get

$$\begin{aligned}\hat{\Psi}_{k+1} &= \hat{\Psi}_k - [\mathcal{L}_\eta\hat{\Psi}_k] + R + \gamma\langle P, \mathcal{L}_\eta\hat{\Psi}_k\rangle_{B,\rho} + \epsilon_k\\ &= kQ_k + Q_0 + R + \gamma\langle P, \mathcal{L}_\eta[kQ_k+Q_0]\\ &\quad - \mathcal{L}_\eta[(k-1)Q_{k-1}+Q_0]\rangle_{B,\rho}\\ &\quad - \mathcal{L}_\eta[kQ_k+Q_0] + \epsilon_k\\ &= kQ_k - kR - \gamma\langle P, \mathcal{L}_\eta[(k-1)Q_{k-1}+Q_0]\rangle_{B,\rho}\\ &\quad - E_{k-1} + (k+1)R + \gamma\langle P, \mathcal{L}[kQ_k+Q_0]\rangle_{B,\rho}\\ &\quad + E_k + Q_0 - \mathcal{L}_\eta[kQ_k+Q_0]\\ &= (k+1)Q_{k+1} + Q_0 - \mathcal{L}_\eta[kQ_k+Q_0]\end{aligned} \tag{30}$$

which proves the desired relation (28). Incidentally, since $\mathcal{L}_\eta q$ is independent of $a \in \mathcal{A}$, a trivial consequence of (28) is that

$$\hat{\pi}_k(a\,|\,b) := \frac{\exp[\eta\{Q_k(b,a)+Q_0(b,a)\}]}{\sum_{a'}\exp[\eta\{Q_k(b,a')+Q_0(b,a')\}]} \tag{31}$$

from which we conclude that

$$\begin{aligned}\mathcal{M}_\eta(kQ_k+Q_0) &= \langle\hat{\pi}_k, kQ_k+Q_0\rangle\\ &= k\langle\hat{\pi}_k, Q_k\rangle + \langle\hat{\pi}_k, Q_0\rangle.\end{aligned} \tag{32}$$

Step 2. Next, we try to explicitly bound $\|Q^{\pi^*}_{B,\rho} - Q_k\|_\infty$. We try to inductively prove the relation

$$\begin{aligned}\|Q^{\pi^*}_{B,\rho} - Q_k\|_\infty &\le \frac{\gamma(4Q_{\max} + \log(|\mathcal{A}|)/\eta)}{(1-\gamma)k}\\ &\quad + \frac{1}{k}\sum_{j=1}^{k}\gamma^{k-j}\|E_{j-1}\|_\infty, \quad \forall k \ge 1.\end{aligned} \tag{33}$$

First we prove the base case for $k = 1$. Our hypothesis $\|Q_0\|_\infty = \|\hat{\Psi}_0\|_\infty \le Q_{\max}$ together with the triangle inequality for norms yield

$$\|Q^{\pi^*}_{B,\rho} - Q_0\|_\infty \le \|Q^{\pi^*}_{B,\rho}\|_\infty + \|Q_0\|_\infty \le 2Q_{\max}. \tag{34}$$

Thus,

$$
\begin{aligned}
&\|Q^{\pi^*}_{B,\rho} - Q_1\|_\infty \\
&= \|\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} - (R + \gamma\langle P, \mathcal{L}_\eta Q_0\rangle_{B,\rho} + E_0)\|_\infty \\
&\le \|\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} - \mathscr{T}^{\pi^*}_{B,\rho} Q_0\|_\infty + \|\mathscr{T}^{\pi^*}_{B,\rho} Q_0 \\
&\quad - (R + \gamma\langle P, \mathcal{L}_\eta Q_0\rangle_{B,\rho})\|_\infty + \|E_0\|_\infty \\
&\le \gamma\|Q^{\pi^*}_{B,\rho} - Q_0\|_\infty + \gamma\Big\|\big\langle P, \mathcal{M}Q_0 - \mathcal{L}_\eta Q_0\big\rangle_{B,\rho}\Big\|_\infty \\
&\quad + \|E_0\|_\infty \\
&\le \gamma\|Q^* - Q_0\|_\infty + \gamma\Big\|\mathcal{M}Q_0 - \mathcal{L}_\eta Q_0\Big\|_\infty + \|E_0\|_\infty \\
&\le \gamma\Big[2Q_{\max} + \frac{\log(|\mathcal{A}|)}{\eta}\Big] + \|E_0\|_\infty
\end{aligned} \tag{35}
$$

where we have made use of (11), Lemma 1 and Proposition 4. This validates the base case of (33).

Now suppose (33) holds up to some $k \ge 1$. Then

$$
\begin{aligned}
&\|Q^{\pi^*}_{B,\rho} - Q_{k+1}\|_\infty \\
&= \Big\|\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} \\
&\quad - \Big(R + \frac{\gamma}{k+1}\langle P, \mathcal{L}_\eta[kQ_k + Q_0]\rangle_{B,\rho} + \frac{E_k}{k+1}\Big)\Big\|_\infty \\
&= \frac{1}{k+1}\Big\|\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} - \mathscr{T}^{\pi^*}_{B,\rho} Q_0 + \mathscr{T}^{\pi^*}_{B,\rho} Q_0 \\
&\quad - \big[(k+1)R + \gamma\big\langle P, \mathcal{L}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho}\big] \\
&\qquad + k(\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} - \mathscr{T}^{\pi^*}_{B,\rho} Q_k + \mathscr{T}^{\pi^*}_{B,\rho} Q_k) - E_k\Big\|_\infty \\
&\le \frac{1}{k+1}\Big[\big\|\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} - \mathscr{T}^{\pi^*}_{B,\rho} Q_0\big\|_\infty \\
&\quad + k\big\|\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} - \mathscr{T}^{\pi^*}_{B,\rho} Q_k\big\|_\infty \\
&\qquad + \Big\|k\mathscr{T}^{\pi^*}_{B,\rho} Q_k + \mathscr{T}^{\pi^*}_{B,\rho} Q_0 \\
&\quad - \big[(k+1)R + \gamma\big\langle P, \mathcal{L}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho}\big]\Big\|_\infty + \|E_k\|_\infty\Big] \\
&\le \frac{1}{k+1}\Big[\gamma\|Q^{\pi^*}_{B,\rho} - Q_0\|_\infty + \gamma k\|Q^{\pi^*}_{B,\rho} - Q_k\|_\infty \\
&\qquad + \Big\|k\mathscr{T}^{\pi^*}_{B,\rho} Q_k + \mathscr{T}^{\pi^*}_{B,\rho} Q_0 \\
&\quad - \big[(k+1)R + \gamma\big\langle P, \mathcal{L}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho}\big]\Big\|_\infty + \|E_k\|_\infty\Big].
\end{aligned} \tag{36}
$$

Now Proposition 5 and Proposition 4 yield

$$
\begin{aligned}
&\big\|k\mathscr{T}^{\pi^*}_{B,\rho} Q_k + \mathscr{T}^{\pi^*}_{B,\rho} Q_0 - (k+1)R \\
&- \gamma\big\langle P, \mathcal{L}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho}\big\|_\infty \\
&\le \big\|\gamma\big\langle P, \mathcal{M}(kQ_k) + \mathcal{M}Q_0 - \mathcal{L}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho}\big\|_\infty \\
&\le \gamma\big\|\mathcal{M}(kQ_k) + \mathcal{M}Q_0 - \mathcal{L}_\eta[kQ_k + Q_0]\big\|_\infty \\
&\le \gamma\big\|\mathcal{M}(kQ_k + Q_0) + 2\mathcal{M}|Q_0| - \mathcal{L}_\eta[kQ_k + Q_0]\big\|_\infty \\
&\le \gamma\Big[2\|Q_0\|_\infty + \frac{\log(|\mathcal{A}|)}{\eta}\Big].
\end{aligned} \tag{37}
$$

Simple computations after substituting (37) into (36) then gives

$$
\begin{aligned}
\|Q^{\pi^*}_{B,\rho} - Q_{k+1}\|_\infty \le \frac{1}{k+1}\Big[&\frac{\gamma(4Q_{\max} + \log(|\mathcal{A}|)/\eta)}{(1-\gamma)} \\
&+ \sum_{j=1}^{k+1}\gamma^{k+1-j}\|E_{j-1}\|_\infty\Big]
\end{aligned} \tag{38}
$$

which verifies the desired relation.

Step 3. We are now ready to prove the main result. The triangle inequality and the contraction property of $\mathscr{T}^{\hat{\pi}_k}$ give

$$
\begin{aligned}
&\|Q^{\pi^*}_{B,\rho} - Q^{\hat{\pi}_k}_{B,\rho}\|_\infty \\
&\le \|Q^{\pi^*}_{B,\rho} - Q_{k+1}\|_\infty + \|Q_{k+1} - \mathscr{T}^{\hat{\pi}_k}_{B,\rho} Q^{\pi^*}_{B,\rho}\|_\infty \\
&\quad + \|\mathscr{T}^{\hat{\pi}_k}_{B,\rho} Q^{\pi^*}_{B,\rho} - \mathscr{T}^{\hat{\pi}_k}_{B,\rho} Q^{\hat{\pi}_k}_{B,\rho}\|_\infty \\
&\le \|Q^{\pi^*}_{B,\rho} - Q_{k+1}\|_\infty + \|Q_{k+1} - \mathscr{T}^{\hat{\pi}_k}_{B,\rho} \\
&\quad + \gamma\|Q^{\pi^*}_{B,\rho} - Q^{\hat{\pi}_k}_{B,\rho}\|_\infty
\end{aligned} \tag{39}
$$

Rearranging, we get

$$
\begin{aligned}
(1-\gamma)\|Q^{\pi^*}_{B,\rho} - Q^{\hat{\pi}_k}_{B,\rho}\|_\infty \le \|Q^{\pi^*}_{B,\rho} - Q_{k+1}\|_\infty& \\
+ \|Q_{k+1} - \mathscr{T}^{\hat{\pi}_k}_{B,\rho} Q^{\pi^*}_{B,\rho}\|_\infty&.
\end{aligned} \tag{40}
$$

Now, we can use (23), (32) and (33) to yield

$$
\begin{aligned}
&\|Q_{k+1} - \mathscr{T}^{\hat{\pi}_k}_{B,\rho} Q^{\pi^*}_{B,\rho}\|_\infty \\
&\le \Big\|R + \frac{\gamma}{k+1}\big\langle P, \mathcal{L}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho} \\
&\quad + \frac{E_k}{k+1} - \mathscr{T}^{\hat{\pi}_k}_{B,\rho} Q^{\pi^*}_{B,\rho}\Big\|_\infty \\
&\le \Big\|\frac{\gamma}{k+1}\big\langle P, \mathcal{L}_\eta[kQ_k + Q_0] - \mathcal{M}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho} \\
&\quad + \frac{\gamma}{k+1}\big\langle P, \mathcal{M}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho} + \frac{E_k}{k+1} \\
&\quad - \big\langle P, \langle\hat{\pi}_k, Q^{\pi^*}_{B,\rho}\rangle\big\rangle_{B,\rho}\Big\|_\infty \\
&\le \frac{\gamma}{k+1}\Big[\frac{\alpha}{1-\gamma} + \sum_{j=1}^{k}\gamma^{k-j}\|E_{j-1}\|_\infty\Big] + \frac{\|E_k\|_\infty}{k+1}
\end{aligned} \tag{41}
$$

where

$$
\alpha := 4Q_{\max} + \frac{\log(|\mathcal{A}|)}{\eta}. \tag{42}
$$

Using the above bound and (33) in (40) and setting $B = \mathcal{R}_{b_0}$ then gives us

$$
\begin{aligned}
(1-\gamma)\|Q^* - Q^{\hat{\pi}_k}\|_\infty \le \frac{2\gamma\alpha}{(1-\gamma)(k+1)}& \\
+ \frac{2}{k+1}\sum_{j=0}^{k}\gamma^{k-j}\|E_j\|_\infty&
\end{aligned} \tag{43}
$$

which completes the proof. □

4 Convergence of the Approximate Scheme

We prove the high-probability loss bound presented in Theorem 2 of the main body. We will need the following result.

Lemma 4. *If* $\|\hat{\Psi}_0\|_\infty \le Q_{\max}$ *then*

$$\|\epsilon_k\|_{B,\infty} \le \frac{4\gamma\log(|\mathcal{A}|)}{\eta(1-\gamma)} + 2Q_{\max} =: U, \quad \forall k \ge 0 \tag{44}$$

for the sequence $(\hat{\Psi}_k)_{k\ge0}$ *generated by the synchronous scheme.*

Proof. For any $(b,a) \in \mathcal{E}_\delta \times \mathcal{A}$, the error for the synchronous scheme is given by

$$\begin{aligned}\epsilon_k(b,a) = R(b,a) - \sum_{i=1}^{N_k(b,a)} \frac{R(s_i,a)}{N_k(b,a)} \\ + \frac{\gamma}{M_k}\sum_{j=1}^{M_k(b,a)} [\mathcal{L}_\eta\hat{\Psi}_k]\big(\tilde{\tau}_{\mathcal{E}_\delta,\rho_1}(b,a,o_k)\big) \\ - \gamma\langle P, [\mathcal{L}_\eta\hat{\Psi}_k]\rangle_{\mathcal{E}_\delta,\rho_1}(b,a).\end{aligned}$$

This and the bound (11) give

$$\begin{aligned}\|\epsilon_k\|_\infty \le 2R_{\max} + \gamma \sup_{(b,a)\in\mathcal{E}_\delta\times\mathcal{A}} \Big|[\mathcal{L}_\eta\hat{\Psi}_k]\big(\tilde{\tau}_{\mathcal{E}_\delta,\rho_1}(b,a,o_k)\big)\Big| \\ + \gamma\big\|\langle P, [\mathcal{L}_\eta\hat{\Psi}_k]\rangle_{\mathcal{E}_\delta,\rho_1}(b,a)\big\|_{\mathcal{E}_\delta,\infty} \\ \le 2R_{\max} + 2\gamma\|\mathcal{L}_\eta\hat{\Psi}_k\|_{\mathcal{E}_\delta,\infty}\end{aligned}$$

so it suffices to bound $\|\mathcal{L}_\eta\hat{\Psi}_k\|_{\mathcal{E}_\delta,\infty}$ for all $k \ge 0$ which we can validate via induction. We claim that

$$\|\mathcal{L}_\eta\hat{\Psi}_k\|_{\mathcal{E}_\delta,\infty} \le \frac{2\log(|\mathcal{A}|)}{\eta(1-\gamma)} + Q_{\max} \quad \forall k \ge 0. \tag{45}$$

The base case follows immediately from (24) so that, for any $b \in \mathcal{B}$,

$$\Big|[\mathcal{L}_\eta\hat{\Psi}_0](b)\Big| \le \log(|\mathcal{A}|)/\eta + \|\hat{\Psi}_0\|_\infty \tag{46}$$

which satisfies the required bound since we hypothesised that $\|\hat{\Psi}_0\|_\infty \le Q_{\max}$. For the induction step, it suffices to fix a $b \in \mathcal{E}_\delta$ and to observe that

$$\begin{aligned}&\Big|[\mathcal{L}_\eta\hat{\Psi}_{k+1}](b)\Big| \\ &= \Big|[\mathcal{L}_\eta\hat{\Psi}_{k+1}](b) - [\mathcal{M}\hat{\Psi}_0](b) + [\mathcal{M}\hat{\Psi}_{k+1}](b)\Big| \\ &\le \Big|[\mathcal{M}\hat{\Psi}_k - [\mathcal{L}_\eta\hat{\Psi}_k] + R \\ &\quad + \gamma[\mathcal{L}_\eta\hat{\Psi}_k](\tilde{\tau}_{\mathcal{E}_\delta,\rho_1}(\cdot,\cdot,o_k))\big](b)\Big| + \frac{\log(|\mathcal{A}|)}{\eta} \\ &\le \frac{\log(|\mathcal{A}|)}{\eta} + \Big|[\mathcal{M}\hat{\Psi}_k](b) - [\mathcal{L}_\eta\hat{\Psi}_k](b)\Big| + R_{\max} \\ &\quad + \gamma\Big|[\mathcal{M}[\mathcal{L}_\eta\hat{\Psi}_k](\tilde{\tau}_{\mathcal{E}_\delta,\rho_1}(\cdot,\cdot,o_k))](b)\Big| \\ &\le \frac{2\log(|\mathcal{A}|)}{\eta} + R_{\max} + \gamma\|\mathcal{L}_\eta\hat{\Psi}_k\|_{\mathcal{E}_\delta,\infty} \\ &\le \frac{2\log(|\mathcal{A}|)}{\eta} + R_{\max} + \frac{2\gamma\log(|\mathcal{A}|)}{\eta(1-\gamma)} + \gamma Q_{\max} \\ &= \frac{2\log(|\mathcal{A}|)}{\eta(1-\gamma)} + Q_{\max}\end{aligned}$$

and the result follows from the arbitrariness of $b \in \mathcal{E}_\delta$. $\square$

Theorem 2 is valid for a *synchronous* backup. In other words, we sample the observations $o_1^{b,a},\dots,o_{N_k}^{b,a}$ from the distribution $P(\cdot\,|\,b,a)$ for *every* $(b,a) \in \mathcal{E}_\delta \times \mathcal{A}$ and then compute $\hat{\Psi}_{k+1}$ according to the synchronous scheme at each iteration k. Let $\mathbf{o}_k := [o_k^{b,a}]_{(b,a)\in\mathcal{E}_\delta\times\mathcal{A}}$ represent the collective sampled random variable after one synchronous iteration.

TODO: Tidy up the discussion about filtrations here. Now, let $(\mathcal{F}_k)_{k\ge0}$ be the filtration generated by the random variables $(\mathbf{o}_i)_{0\le i\le k}$. Intuitively, each $\mathcal{F}_k$ can be seen as the set of events that can be distinguished as true or false after having observed $(\mathbf{o}_i)_{0\le i\le k}$. By our definition of the approximate sequence, it is clear that

$$\mathbb{E}[\epsilon_k(b,a)\,|\,\mathcal{F}_{k-1}] = 0 \quad \forall b, \forall a, \forall k \ge 1 \tag{47}$$

from which we can conclude that $E_k(b,a)$ is a martingale with respect to $(\mathcal{F}_k)_{k\ge0}$ satisfying $E_0(b,a) = 0$. Hence, we can apply Theorem 1 with the uniform bound from Lemma 4 to conclude that, for any $\beta > 0$,

$$\begin{aligned}&\mathbb{P}\Big(\sup_{0\le j\le k}\|E_j\|_{\mathcal{E}_\delta,\infty} \ge \beta\Big) \\ &= \mathbb{P}\Big(\sup_{(b,a)\in\mathcal{E}_\delta\times\mathcal{A}}\sup_{0\le j\le k}|E_j(b,a)| \ge \beta\Big) \\ &= \mathbb{P}\Big(\bigcup_{(b,a)\in\mathcal{E}_\delta\times\mathcal{A}}\Big\{\sup_{0\le j\le k}|E_j(b,a)| \ge \beta\Big\}\Big) \\ &\le \sum_{(b,a)\in\mathcal{E}_\delta\times\mathcal{A}}\mathbb{P}\Big(\sup_{0\le j\le k}|E_j(b,a)| \ge \beta\Big) \\ &= 2|\mathcal{E}_\delta||\mathcal{A}|\exp\Big(-\frac{2\beta^2}{(k+1)U^2}\Big).\end{aligned}$$

where U is the uniform error bound obtained in (44). Hence

$$\mathbb{P}\Big(\sup_{0\le j\le k}\|E_j\|_{\mathcal{E}_\delta,\infty} < \beta\Big) \ge 1 - 2|\mathcal{E}_\delta||\mathcal{A}|\exp\Big[-\frac{2\beta^2}{(k+1)U^2}\Big] =: 1-\alpha$$

and with probability at least $1-\alpha$ we have

$$\begin{aligned}&\sum_{j=0}^{k}\gamma^{k-j}\|E_j\|_{\mathcal{E}_\delta,\infty} \le \sum_{j=0}^{k}\gamma^{k-j}\sup_{0\le j\le k}\|E_j\|_{\mathcal{E}_\delta,\infty} \\ &\le (1-\gamma)^{-1}\sup_{0\le j\le k}\|E_j\|_{\mathcal{E}_\delta,\infty} \le \frac{U}{1-\gamma}\sqrt{\frac{k+1}{2}\log\Big[\frac{2|\mathcal{E}_\delta||\mathcal{A}|}{\alpha}\Big]} \\ &\le \frac{4\gamma\alpha}{1-\gamma}\sqrt{\frac{k+1}{2}\log\Big[\frac{2|\mathcal{E}_\delta||\mathcal{A}|}{\alpha}\Big]}.\end{aligned}$$

where $\alpha := 4Q_{\max} + \log(|\mathcal{A}|)/\eta$. Finally, we conclude from (**??**) that

$$\|Q^* - Q^{\hat{\pi}}_{\mathcal{E}_\delta,\rho_1}\|_\infty \leq \frac{2}{(1-\gamma)(k+1)}\left[\frac{\gamma\alpha}{1-\gamma} + \sum_{j=0}^{k}\gamma^{k-j}\|E_j\|_{\mathcal{E}_\delta,\infty}\right] \leq \frac{2\gamma B}{(1-\gamma)^2}\left[\frac{1}{k+1} + \frac{1}{1-\gamma}\sqrt{8\log\left[\frac{2|\mathcal{E}_\delta||\mathcal{A}|}{\alpha}\right]}\frac{1}{\sqrt{k+1}}\right] + \frac{\gamma\delta Q_{\max}}{(1-\gamma)}.$$

which concludes the proof.

4.1 A Maximal Azuma-Hoeffding Inequality

We employ a maximal version of the Azuma-Hoeffding inequality (see e.g. [Cesa-Bianchi and Lugosi, 2006]). It follows by replacing Markov's inequality with the Doob's maximal inequality for sub- or supermartingales (see [Doob, 1953] p. 314) in the proof of the standard (i.e non-maximal) version of the inequality (see e.g. [Hoeffding, 1963]). Intuitively, it bounds the likelihood of a martingale (or, more generally, a submartingale) having ever exceeded a given distance from its starting point, where the bound increases to one with the number of steps. As such, it can be seen as a concentration bound.

Theorem 1 (Maximal Azuma-Hoeffding Inequality)**.** *Let $(M_t)_{t\geq 0}$ be a discrete-time martingale with respect to a given filtration $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ on an arbitrary probability space $(\Omega, \mathcal{F}_\infty, \mathbb{P})$. Assume that there are $\mathbb{F}$-predictable processes processes $(A_t)_{t\geq 0}$ and $(B_t)_{t\geq 0}$ and constants $0 < c_t < +\infty$ such that:*

$$A_t \leq M_t - M_{t-1} \leq B_t \quad \textit{and} \quad B_t - A_t \leq c_t \quad P\textit{-a.s.}. \tag{48}$$

Then for all $\beta > 0$

$$\mathbb{P}\left[\sup_{0\leq s\leq t}(M_s - M_0) \geq \beta\right] \leq \exp\left(-\frac{2\beta^2}{\sum_{0\leq s\leq t} c_s^2}\right). \tag{49}$$

References

[Bertsekas, 2008] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, 3rd edition, 2008.

[Cesa-Bianchi and Lugosi, 2006] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[Doob, 1953] J. L. Doob. *Stochastic Processes*. Wiley, 1953.

[Hoeffding, 1963] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[Kolmogorov and Fomin, 1970] A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis*. Dover, 1970.

[Lee *et al.*, 2007] Wee Lee, Nan Rong, and David Hsu. What makes some POMDP problems easy to approximate? In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[MacKay, 2003] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[Ross, 1970] S. M. Ross. *Applied Probability Models with Optimization Applications*. Dover, 1970.