Appendix: On Linear Convergence of Policy Gradient Methods for Finite MDPs

A Proof of supporting lemmas

We give proofs of Lemmas 2 and 3, which were excluded from the main text.

Lemma 2. *For any state s,* $\frac{\pi^{t+1}(s,i)}{\pi^t(s,i)} \leq \frac{1}{2} \;\; \forall i \in O_t^-(s)$.

Proof. The proof follows a simple argument. By definition, for any $i \in O_t^-(s)$:

$$\begin{aligned} \big(Q^t(s,i) - Q^t(s,1)\big) &\geq \frac{\epsilon}{c} \\ \Rightarrow \; \alpha_t(s)\big(Q^t(s,i) - Q^t(s,1)\big) &\geq \log \frac{2}{\pi^t(s,1)} \end{aligned}$$

which follows by the definition, $\alpha_t(s) \geq \frac{c}{\epsilon} \log \frac{2}{\pi^t(s,1)}$ which implies $\frac{\epsilon}{c} \geq \frac{1}{\alpha_t(s)} \log \frac{2}{\pi^t(s,1)}$. Rearranging, we get

$$\log\Big(\pi^t(s,1)e^{-\alpha_t(s)Q^t(s,1)}\Big) + \log\Big(\frac{1}{2}\Big) \geq -\alpha_t(s)Q^t(s,i)$$

Define, $Z_t = \Big(\sum_{j=1}^k \pi^t(s,j)e^{-\alpha_t(s)Q^t(s,j)}\Big)$. Then,

$$\log(Z_t) \geq \log\Big(\pi^t(s,1)e^{-\alpha_t(s)Q^t(s,1)}\Big)$$

which holds as all the terms in Z_t are positive, i.e. $\pi^t(s,j)e^{-\alpha_t(s)Q^t(s,j)} > 0 \;\; \forall j \in \{1,2,\ldots,k\}$, and $\log(\cdot)$ is a monotonic transformation. Rearranging, we get our desired result.

$$\begin{aligned} \log\Big(\frac{Z_t}{2}\Big) \geq \log\Big(\frac{\pi^t(s,1)}{2}e^{-\alpha_t(s)Q^t(s,1)}\Big) &\geq -\alpha_t(s)Q^t(s,i) \\ \Rightarrow \; \frac{\pi^{t+1}(s,i)}{\pi^t(s,i)} = \frac{1}{Z_t}e^{-\alpha_t(s)Q^t(s,i)} &\leq \frac{1}{2}. \end{aligned}$$

□

Lemma 3 (Progress quantification)**.** *Let $J_{\pi^t}(s)$ denote the cost-to-go function for policy π^t from any starting state $s \in \mathcal{S}$. Then,*

$$T_{\pi^{t+1}}J_{\pi^t}(s) - J_{\pi^t}(s) \leq \frac{1}{2} \cdot \big(TJ_{\pi^t}(s) - J_{\pi^t}(s)\big) + \frac{\epsilon}{c}$$

Proof. Fix any state $s \in \mathcal{S}$. Without loss of generality, we assume the following ordering on Q-values: $Q^t(s,1) < Q^t(s,2) \ldots < Q^t(s,k)$ which implies that the policy iteration update, π_t^+ puts the entire mass on action 1, which is the best action under the current policy π^t. That is, $\pi_+^t(s,1)=1$ and $\pi_+^t(s,i)=0\ \forall i\neq 1$. Consider,

$$\begin{aligned}
T_{\pi^{t+1}}J_{\pi^t}(s)-TJ_{\pi^t}(s) &= \langle \pi^{t+1}(s,\cdot)-\pi_+^t(s,\cdot), Q^t(s,\cdot)\rangle \\
&= (\pi^{t+1}(s,1)-1)Q^t(s,1)+\sum_{j=2}^{k}\pi^{t+1}(s,j)Q^t(s,j) \\
&= -\sum_{j=2}^{k}\pi^{t+1}(s,j)Q^t(s,1)+\sum_{j=2}^{k}\pi^{t+1}(s,j)Q^t(s,j) \\
&= \sum_{j=2}^{k}\pi^{t+1}(s,j)\big(Q^t(s,j)-Q^t(s,1)\big) \\
&= \sum_{j\in\mathcal{O}_t^-}\pi^{t+1}(s,j)\big(Q^t(s,j)-Q^t(s,1)\big)+\sum_{j\in\mathcal{O}_t^+}\pi^{t+1}(s,j)\big(Q^t(s,j)-Q^t(s,1)\big) \\
&= \sum_{j\in\mathcal{O}_t^-}\frac{\pi^{t+1}(s,j)}{\pi^t(s,j)}\pi^t(s,j)\big(Q^t(s,j)-Q^t(s,1)\big)+\sum_{j\in\mathcal{O}_t^+}\pi^{t+1}(s,j)\underbrace{\big(Q^t(s,j)-Q^t(s,1)\big)}_{<\frac{\epsilon}{c}} \\
&\le \frac{1}{2}\sum_{j\in\mathcal{O}_t^-}\pi^t(s,j)\big(Q^t(s,j)-Q^t(s,1)\big)+\frac{\epsilon}{c} \\
&\le \frac{1}{2}\left(\sum_{j=2}^{k}\pi^t(s,j)(Q^t(s,j)-Q^t(s,1))\right)+\frac{\epsilon}{c} \\
&= \frac{1}{2}\left(\sum_{j=2}^{k}\pi^t(s,j)Q^t(s,j)-\sum_{j=2}^{k}\pi^t(s,j)Q^t(s,1)\right)+\frac{\epsilon}{c} \\
&= \frac{1}{2}\left(\big(\pi^t(s,1)-1\big)Q^t(s,1)+\sum_{j=2}^{k}\pi^t(s,j)Q^t(s,j)\right)+\frac{\epsilon}{c} \\
&= \frac{1}{2}\langle \pi^t(s,\cdot)-\pi_+^t(s,\cdot), Q^t(s,\cdot)\rangle+\frac{\epsilon}{c} \\
&= \frac{1}{2}\left(J_{\pi^t}(s)-TJ_{\pi^t}(s)\right)+\frac{\epsilon}{c} \qquad (13)
\end{aligned}$$

where we used that $\frac{\pi^{t+1}(s,j)}{\pi^t(s,j)}\le\frac{1}{2}\ \forall j\in\mathcal{O}_t^-(s)$ as shown above in Lemma 2 along with the fact that $(Q^t(s,j)-Q^t(s,1))\le\frac{\epsilon}{c}\ \forall j\in\mathcal{O}_t^+(s)$, which follows by definition. Subtracting $J_{\pi^t}(s)$ from both sides in (13) and rearranging terms gives our desired result,

$$T_{\pi^{t+1}}J_{\pi^t}(s)-J_{\pi^t}(s)\le\frac{1}{2}\cdot\left(TJ_{\pi^t}(s)-J_{\pi^t}(s)\right)+\frac{\epsilon}{c}.$$

□

B Details of MDP in Figure 1

We used the following two state three action MDP, $P\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$, $g\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\gamma,\rho\in\mathbb{R}^{|\mathcal{S}|}$, to generate Figure 1.

$$P=\begin{bmatrix}0.666066 & 0.333934\\ 0.662211 & 0.337789\\ 0.441947 & 0.558053\\ 0.391257 & 0.608743\\ 0.452186 & 0.547814\\ 0.035519 & 0.964481\end{bmatrix},\ g=\begin{bmatrix}0.079718\\ 0.629733\\ 0.717644\\ 0.673362\\ 0.762623\\ 0.541251\end{bmatrix},\ \gamma=0.9,\ \rho=\begin{bmatrix}0.168831\\ 0.831169\end{bmatrix}$$

Policy π for the two states s_1 and s_2 was taken to be,

$$\pi(s_1) = \begin{bmatrix} 0.449416 \\ 0.251788 \\ 0.298796 \end{bmatrix}, \pi(s_2) = \begin{bmatrix} 0.318626 \\ 0.346284 \\ 0.335090 \end{bmatrix}.$$