Appendix

A. Preliminaries

A.1. Derivation of entropy-regularized NPG methods

This subsection establishes the equivalence between the update rules (15) and (18). Such derivations are inherently similar to the ones for the NPG update rule (without entropy regularization) (see, e.g., Agarwal et al. (2019)); we provide the proof here for pedagogical reasons.

First of all, let us follow the convention to introduce the advantage function $A_\tau^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of a policy π w.r.t. the entropy-regularized value function:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad A_\tau^\pi(s,a) := Q_\tau^\pi(s,a) - \tau \log \pi(a|s) - V_\tau^\pi(s) \tag{A.1}$$

with Q_τ^π defined in (11a), which reflects the gain one can harvest by executing action a instead of following the policy π in state s. This advantage function plays a crucial role in the calculation of policy gradients, due to the following fundamental relation (see Appendix C.6 for the proof):

Lemma 1 *Under softmax parameterization (7), the gradient of the regularized value function satisfies*

$$\frac{\partial V_\tau^{\pi_\theta}(\rho)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A_\tau^{\pi_\theta}(s,a); \tag{A.2a}$$

$$\left[\left(\mathcal{F}_\rho^\theta\right)^\dagger \nabla_\theta V_\tau^{\pi_\theta}(\rho) \right](s,a) = \frac{1}{1-\gamma} A_\tau^{\pi_\theta}(s,a) + c(s) \tag{A.2b}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, where $c(s) := \sum_a \pi_\theta(a|s) w_{s,a}$ is some function depending only on s.

It is worth highlighting that the search direction of NPG, given in (A.2b), is invariant to the choice of ρ. With the above calculations in place, it is seen that for any $s \in \mathcal{S}$, the regularized NPG update rule (15) results in a policy update as follows

$$\begin{aligned}
\pi^{(t+1)}(a|s) &\overset{\text{(i)}}{\propto} \exp\left(\theta^{(t+1)}(s,a)\right) \overset{\text{(ii)}}{=} \exp\left(\theta^{(t)}(s,a) + \eta \left[\left(\mathcal{F}_\rho^{\theta^{(t)}}\right)^\dagger \nabla_\theta V_\tau^{(t)}(\rho) \right](s,a)\right) \\
&\overset{\text{(iii)}}{\propto} \exp\left(\theta^{(t)}(s,a) + \frac{\eta}{1-\gamma} A_\tau^{(t)}(s,a)\right) \\
&\overset{\text{(iv)}}{\propto} \pi^{(t)}(a|s) \exp\left(\frac{\eta}{1-\gamma} Q_\tau^{(t)}(s,a) - \frac{\eta\tau}{1-\gamma} \log \pi^{(t)}(a|s)\right) \\
&= \left(\pi^{(t)}(a|s)\right)^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta}{1-\gamma} Q_\tau^{(t)}(s,a)\right).
\end{aligned}$$

where we use $A_\tau^{(t)}$ to abbreviate $A_\tau^{\pi^{(t)}}$. Here, (i) uses the definition of the softmax policy, (ii) comes from the update rule (15), (iii) is a consequence of (A.2b) (since $c(\cdot)$ does not depend on a), whereas (iv) results from the definition (A.1) and the fact that $V_\tau^\pi(\cdot)$ is not dependent on a. This validates the equivalence between (15) and (18).

A.2. Basic facts about the function $\log(\|\exp(\theta)\|_1)$

In the current paper, we often encounter the function $\log\big(\|\exp(\theta)\|_1\big) := \log\big(\sum_{1\le a\le|\mathcal{A}|}\exp(\theta_a)\big)$ for any vector $\theta=[\theta_a]_{1\le a\le|\mathcal{A}|}\in\mathbb{R}^{|\mathcal{A}|}$. To facilitate analysis, we single out several basic properties concerning this function, which will be used multiple times when establishing our main results. For notational convenience, we denote by $\pi_\theta\in\mathbb{R}^{|\mathcal{A}|}$ the softmax transform of θ such that

$$\pi_\theta(a)=\frac{\exp(\theta_a)}{\sum_{1\le j\le|\mathcal{A}|}\exp(\theta_j)},\qquad 1\le a\le|\mathcal{A}|. \tag{A.3}$$

By straightforward calculations, the gradient of the function $\log\big(\|\exp(\theta)\|_1\big)$ is given by

$$\nabla_\theta\log\big(\|\exp(\theta)\|_1\big)=\frac{1}{\|\exp(\theta)\|_1}\exp(\theta)=\pi_\theta;. \tag{A.4}$$

Difference of log policies. In the analysis, we often need to control the difference of two policies, towards which the following bounds prove useful. To begin with, the mean value theorem reveals a Lipschitz continuity property (w.r.t. the ℓ_∞ norm): for any $\theta_1,\theta_2\in\mathbb{R}^{|\mathcal{A}|}$,

$$\begin{aligned}\big|\log\big(\|\exp(\theta_1)\|_1\big)-\log\big(\|\exp(\theta_2)\|_1\big)\big| &= \big|\big\langle\theta_1-\theta_2,\nabla_\theta\log\big(\|\exp(\theta)\|_1\big)|_{\theta=\theta_c}\big\rangle\big|\\ &\le\|\theta_1-\theta_2\|_\infty\big\|\nabla_\theta\log\big(\|\exp(\theta)\|_1\big)|_{\theta=\theta_c}\big\|_1=\|\theta_1-\theta_2\|_\infty,\end{aligned} \tag{A.5}$$

where θ_c is a certain convex combination of θ_1 and θ_2, and the second line relies on (A.4). In addition, for any two vectors π_{θ_1} and π_{θ_2} defined w.r.t. $\theta_1,\theta_2\in\mathbb{R}^{|\mathcal{A}|}$ (see (A.3)), one has

$$\|\log\pi_{\theta_1}-\log\pi_{\theta_2}\|_\infty\le 2\|\theta_1-\theta_2\|_\infty, \tag{A.6}$$

where $\log(\cdot)$ denotes entrywise operation. To justify (A.6), we observe from the definition (A.3) that

$$\|\log\pi_{\theta_1}-\log\pi_{\theta_2}\|_\infty\le\|\theta_1-\theta_2\|_\infty+\Big|\log\big(\|\exp(\theta_1)\|_1\big)-\log\big(\|\exp(\theta_2)\|_1\big)\Big|\le 2\|\theta_1-\theta_2\|_\infty,$$

where the last inequality is a consequence of (A.5).

B. Proof for the bandit case (Proposition 1)

We start by defining an auxiliary sequence $\xi^{(t)}\in\mathbb{R}^{|\mathcal{A}|}$ $(t\ge 0)$ recursively as follows

$$\begin{aligned}\xi^{(0)}&:=\|\exp(r/\tau)\|_1\cdot\pi^{(0)},\\ \xi^{(t+1)}(a)&:=\big(\xi^{(t)}(a)\big)^{1-\tau\eta}\exp\big(\eta r(a)\big),\qquad a\in\mathcal{A}.\end{aligned}$$

When combined with (21), it is easily seen that $\pi^{(t)}(\cdot)\propto\xi^{(t)}(\cdot)$ and, as a result, $\pi^{(t)}=\xi^{(t)}/\big\|\xi^{(t)}\big\|_1$.

By construction, the auxiliary sequence satisfies the following property

$$\log\big(\xi^{(t+1)}(a)\big)-r(a)/\tau=(1-\tau\eta)\log\big(\xi^{(t)}(a)\big)+\eta r(a)-r(a)/\tau$$

$$= (1-\tau\eta)\left(\log\left(\xi^{(t)}(a)\right) - r(a)/\tau\right),$$

thus indicating that

$$\left\|\log\xi^{(t)} - r/\tau\right\|_\infty \le (1-\tau\eta)^t \left\|\log\xi^{(0)} - r/\tau\right\|_\infty. \tag{A.7}$$

This taken together with the optimal policy $\pi_\tau^\star = \mathsf{softmax}(r/\tau) \propto \exp(r/\tau)$ leads to

$$\begin{aligned}\left\|\log\pi^{(t)} - \log\pi_\tau^\star\right\|_\infty \le 2\left\|\log\xi^{(t)} - r/\tau\right\|_\infty &\le 2(1-\tau\eta)^t \left\|\log\xi^{(0)} - r/\tau\right\|_\infty \\ &= 2(1-\tau\eta)^t \left\|\log\pi^{(0)} + \left(\log\|\exp(r/\tau)\|_1\right)\cdot\mathbf{1} - r/\tau\right\|_\infty \\ &= 2(1-\tau\eta)^t \left\|\log\pi^{(0)} - \log\pi_\tau^\star\right\|_\infty,\end{aligned}$$

where the first line follows from the inequality (A.6), the second line follows from the expression (A.7), whereas the last line follows from the form of $\pi_\tau^\star$. We have thus completed the proof of Proposition 1.

C. Proof for key lemmas

C.1. Proof of Lemma 1

To begin with, the regularized NPG update rule (see (18) in Algorithm 1) indicates that

$$\log\pi^{(t+1)}(a|s) = \left(1 - \frac{\eta\tau}{1-\gamma}\right)\log\pi^{(t)}(a|s) + \frac{\eta}{1-\gamma}Q_\tau^{(t)}(s,a) - \log Z^{(t)}(s), \tag{A.8}$$

where $Z^{(t)}$ is some quantity depending only on the state s (but not the action a). Rearranging terms gives

$$-\tau\log\pi^{(t)}(a|s) + Q_\tau^{(t)}(s,a) = \frac{1-\gamma}{\eta}\left(\log\pi^{(t+1)}(a|s) - \log\pi^{(t)}(a|s)\right) + \frac{1-\gamma}{\eta}\log Z^{(t)}(s). \tag{A.9}$$

This in turn allows us to express $V_\tau^{(t)}(s_0)$ for any $s_0 \in \mathcal{S}$ as follows

$$\begin{aligned}V_\tau^{(t)}(s_0) &= \mathop{\mathbb{E}}_{a_0\sim\pi^{(t)}(\cdot|s_0)}\left[-\tau\log\pi^{(t)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0)\right] \\ &= \mathop{\mathbb{E}}_{a_0\sim\pi^{(t)}(\cdot|s_0)}\left[\frac{1-\gamma}{\eta}\log Z^{(t)}(s_0)\right] + \mathop{\mathbb{E}}_{a_0\sim\pi^{(t)}(\cdot|s_0)}\left[\frac{1-\gamma}{\eta}\left(\log\pi^{(t+1)}(a_0|s_0) - \log\pi^{(t)}(a_0|s_0)\right)\right] \\ &= \frac{1-\gamma}{\eta}\log Z^{(t)}(s_0) - \frac{1-\gamma}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\|\,\pi^{(t+1)}(\cdot|s_0)\right) \\ &= \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[\frac{1-\gamma}{\eta}\log Z^{(t)}(s_0)\right] - \frac{1-\gamma}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\|\,\pi^{(t+1)}(\cdot|s_0)\right),\end{aligned} \tag{A.10}$$

where the first identity makes use of the definitions (8) and (11a), the second line follows from (A.9), the third line relies on the definition of the KL divergence, and the last line follows since $Z^{(t)}(s)$ does not depend on a. Invoking (A.9) again to rewrite $\log Z^{(t)}(s_0)$ appearing in the first term of (A.10), we reach

$$V_\tau^{(t)}(s_0) = \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[-\tau\log\pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0) + \left(\tau - \frac{1-\gamma}{\eta}\right)\left(\log\pi^{(t+1)}(a_0|s_0) - \log\pi^{(t)}(a_0|s_0)\right)\right]$$

$$
\begin{aligned}
&\quad - \frac{1-\gamma}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\big\|\,\pi^{(t+1)}(\cdot|s_0)\right)\\
&= \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[-\tau\log\pi^{(t+1)}(a_0|s_0)+Q_\tau^{(t)}(s_0,a_0)\right]+\left(\tau-\frac{1-\gamma}{\eta}\right)\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0)\,\big\|\,\pi^{(t)}(\cdot|s_0)\right)\\
&\quad - \frac{1-\gamma}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\big\|\,\pi^{(t+1)}(\cdot|s_0)\right)\\
&= \mathop{\mathbb{E}}_{\substack{a_0\sim\pi^{(t+1)}(\cdot|s_0),\\ s_1\sim P(\cdot|s_0,a_0)}}\left[-\tau\log\pi^{(t+1)}(a_0|s_0)+r(s_0,a_0)+\gamma V_\tau^{(t)}(s_1)\right]\\
&\qquad -\left(\frac{1-\gamma}{\eta}-\tau\right)\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0)\,\big\|\,\pi^{(t)}(\cdot|s_0)\right)-\frac{1-\gamma}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\big\|\,\pi^{(t+1)}(\cdot|s_0)\right),
\end{aligned}
\tag{A.11}
$$

where the second line uses the definition of the KL divergence, and the third line expands $Q_\tau^{(t)}$ using the definition (11a).

To finish up, applying the above relation (A.11) recursively to expand $V_\tau^{(t)}(s_i)$ $(i\geq 1)$, we arrive at

$$
\begin{aligned}
V_\tau^{(t)}(s_0) &= \mathop{\mathbb{E}}_{\substack{a_i\sim\pi^{(t+1)}(\cdot|s_i),\\ s_{i+1}\sim P(\cdot|s_i,a_i),\forall i\geq 0}}\Bigg[\sum_{i=0}^{\infty}\gamma^i\left\{r(s_i,a_i)-\tau\log\pi^{(t+1)}(a_i|s_i)\right\}\\
&\qquad -\sum_{i=0}^{\infty}\gamma^i\left\{\left(\frac{1-\gamma}{\eta}-\tau\right)\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_i)\,\big\|\,\pi^{(t)}(\cdot|s_i)\right)+\frac{1-\gamma}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s_i)\,\big\|\,\pi^{(t+1)}(\cdot|s_i)\right)\right\}\Bigg]\\
&= V_\tau^{(t+1)}(s_0)-\mathop{\mathbb{E}}_{s\sim d_{s_0}^{(t+1)}}\left[\left(\frac{1}{\eta}-\frac{\tau}{1-\gamma}\right)\mathsf{KL}(\pi^{(t+1)}(\cdot|s)\,\big\|\,\pi^{(t)}(\cdot|s))+\frac{1}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\big\|\,\pi^{(t+1)}(\cdot|s)\right)\right],
\end{aligned}
\tag{A.12}
$$

where the second line follows since the regularized value function $V_\tau^{(t+1)}$ can be viewed as the value function of $\pi^{(t+1)}$ with adjusted rewards $r_\tau^{(t+1)}(s,a):=r(s,a)-\tau\log\pi^{(t+1)}(a|s)$. Averaging the initial state s_0 over the distribution ρ concludes the proof.

C.2. Proof of Lemma 2

In the sequel, we prove each claim in Lemma 2 in order.

Proof of Eqn. (36). Jensen's inequality tells us that: for any $s\in\mathcal{S}$ one has

$$
\begin{aligned}
\mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}\Big[Q(s,a)-\tau\log\pi(a|s)\Big] &= \tau\sum_a\pi(a|s)\log\left(\frac{\exp\left(Q(s,a)/\tau\right)}{\pi(a|s)}\right)\\
&\leq \tau\log\left(\sum_a\pi(a|s)\frac{\exp\left(Q(s,a)/\tau\right)}{\pi(a|s)}\right)\\
&= \tau\log\left(\sum_a\exp\left(Q(s,a)/\tau\right)\right)=\tau\log\left(\left\|\exp\left(Q(s,\cdot)/\tau\right)\right\|_1\right),
\end{aligned}
\tag{A.13}
$$

where in the second line, equality is attained if $\pi(\cdot|s)\propto\exp(Q(s,\cdot)/\tau)$. This immediately gives rise to

$$
\mathcal{T}_\tau(Q)(s,a)=r(s,a)+\gamma\mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[\max_{\pi(\cdot|s')\in\Delta(\mathcal{A})}\mathop{\mathbb{E}}_{a'\sim\pi(\cdot|s')}\Big[Q(s',a')-\tau\log\pi(a'|s')\Big]\right]
$$

$$= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \Big[\tau \log\big(\big\|\exp\big(Q(s',\cdot)/\tau\big)\big\|_1\big)\Big].$$

Proof of Eqn. (37). Recall the characterization of $\pi_\tau^\star$ and $V_\tau^\star$ established in Nachum et al. (2017):

$$\pi_\tau^\star(a|s) = \exp\left(\frac{Q_\tau^\star(s,a) - V_\tau^\star(s)}{\tau}\right), \tag{A.14a}$$

$$V_\tau^\star(s) = \tau \log\big(\big\|\exp\big(Q_\tau^\star(s,\cdot)/\tau\big)\big\|_1\big). \tag{A.14b}$$

Substitution into the expression (36) tells us that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
\mathcal{T}_\tau\big(Q_\tau^\star\big)(s,a) &= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \Big[\tau \log\big(\big\|\exp\big(Q_\tau^\star(s',\cdot)/\tau\big)\big\|_1\big)\Big] \\
&= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \big[V_\tau^\star(s')\big] \\
&= Q_\tau^\star(s,a),
\end{aligned}$$

where the second line results from (A.14b), and the last line follows from the definition of the soft Q-function.

Proof of Eqn. (38). Invoking again the expression (36), we can demonstrate that for any Q_1 and Q_2,

$$\begin{aligned}
&\Big|\mathcal{T}_\tau(Q_1)(s,a) - \mathcal{T}_\tau(Q_2)(s,a)\Big| \\
&\quad= \left|\gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \Big[\tau \log\big(\big\|\exp\big(Q_1(s',\cdot)/\tau\big)\big\|_1\big)\Big] - \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \Big[\tau \log\big(\big\|\exp\big(Q_2(s',\cdot)/\tau\big)\big\|_1\big)\Big]\right| \\
&\quad= \gamma\tau \left|\mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \Big[\log\big(\big\|\exp\big(Q_1(s',\cdot)/\tau\big)\big\|_1\big) - \log\big(\big\|\exp\big(Q_2(s',\cdot)/\tau\big)\big\|_1\big)\Big]\right| \\
&\quad\le \gamma\tau \left\|Q_1/\tau - Q_2/\tau\right\|_\infty \\
&\quad= \gamma \left\|Q_1 - Q_2\right\|_\infty
\end{aligned}$$

holds for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, where the inequality follows from the Lipschitz property (A.5).

C.3. Proof of Lemma 3

For any state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$, we observe that

$$\begin{aligned}
&Q_\tau^\star(s,a) - Q_\tau^{(t+1)}(s,a) \\
&\quad= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \big[V_\tau^\star(s')\big] - \Big(r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \big[V_\tau^{(t+1)}(s')\big]\Big) \\
&\quad= \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \left[\tau \log\left(\left\|\exp\left(\frac{Q_\tau^\star(s',\cdot)}{\tau}\right)\right\|_1\right)\right] - \gamma \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a),\\ a'\sim\pi^{(t+1)}(\cdot|s')}} \Big[Q_\tau^{(t+1)}(s',a') - \tau\log\pi^{(t+1)}(a'|s')\Big],
\end{aligned} \tag{A.15}$$

where the first step invokes the definition (11a) of Q_τ, and the second step is due to the expression (A.14b) of $V_\tau^\star$. To continue, recall that $\pi^{(t)}$ is related to $\xi^{(t)}$ as

$$\forall s \in \mathcal{S}: \qquad \pi^{(t)}(\cdot|s) = \frac{1}{\|\xi^{(t)}(s,\cdot)\|_1}\xi^{(t)}(s,\cdot) \tag{A.16}$$

which can be seen by comparing (42) with (18). This in turn leads to

$$\begin{aligned}\log \pi^{(t+1)}(a|s) &= \log \xi^{(t+1)}(s,a) - \log\left(\left\|\xi^{(t+1)}(s,\cdot)\right\|_1\right) \\ &= \alpha \log \xi^{(t)}(s,a) + (1-\alpha)\frac{Q_\tau^{(t)}(s,a)}{\tau} - \log\left(\left\|\xi^{(t+1)}(s,\cdot)\right\|_1\right),\end{aligned} \tag{A.17}$$

where the second line comes from (42b). By plugging (A.17) into (A.15) we obtain

$$\begin{aligned}Q_\tau^\star(s,a) - Q_\tau^{(t+1)}(s,a) = \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[\tau\log\left(\left\|\exp\left(Q_\tau^\star(s',\cdot)/\tau\right)\right\|_1\right) - \tau\log\left(\left\|\xi^{(t+1)}(s',\cdot)\right\|_1\right)\right] \\ - \gamma \mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a),\\ a'\sim\pi^{(t+1)}(\cdot|s')}}\left[Q_\tau^{(t+1)}(s',a') - \tau\underbrace{\left(\alpha\log\xi^{(t)}(s',a') + (1-\alpha)\frac{Q_\tau^{(t)}(s',a')}{\tau}\right)}_{=\log\xi^{(t+1)}(s',a')}\right]\end{aligned} \tag{A.18}$$

for any $(s,a)\in\mathcal{S}\times\mathcal{A}$. In the sequel, we bound each term on the right-hand side of (A.18) separately.

- In view of the property (A.5), the first term on the right-hand side of (A.18) can be bounded by

$$\tau\log\left(\left\|\exp\left(Q_\tau^\star(s',\cdot)/\tau\right)\right\|_1\right) - \tau\log\left(\left\|\xi^{(t+1)}(s',\cdot)\right\|_1\right) \le \left\|Q_\tau^\star - \tau\log\xi^{(t+1)}\right\|_\infty.$$

- Regarding the second term, the monotonicity (33) of the soft Q-function allows us to derive

$$\begin{aligned}&Q_\tau^{(t+1)}(s,a) - \tau\left(\alpha\log\xi^{(t)}(s,a) + (1-\alpha)Q_\tau^{(t)}(s,a)/\tau\right) \\ &\quad\ge Q_\tau^{(t)}(s,a) - \tau\left(\alpha\log\xi^{(t)}(s,a) + (1-\alpha)Q_\tau^{(t)}(s,a)/\tau\right) \\ &\quad= \alpha\left(Q_\tau^{(t)}(s,a) - \tau\log\xi^{(t)}(s,a)\right) \\ &\quad\overset{\text{(i)}}{=} \alpha\left(\alpha\left(Q_\tau^{(t-1)}(s,a) - \tau\log\xi^{(t-1)}(s,a)\right) + Q_\tau^{(t)}(s,a) - Q_\tau^{(t-1)}(s,a)\right) \\ &\quad\overset{\text{(ii)}}{\ge} \alpha^2\left(Q_\tau^{(t-1)}(s,a) - \tau\log\xi^{(t-1)}(s,a)\right) \\ &\quad\overset{\text{(iii)}}{\ge} \alpha^{t+1}\left(Q_\tau^{(0)}(s,a) - \tau\log\xi^{(0)}(s,a)\right) \\ &\quad\overset{\text{(iv)}}{\ge} -\alpha^{t+1}\left\|Q_\tau^{(0)} - \tau\log\xi^{(0)}\right\|_\infty\end{aligned}$$

for any $(s,a)\in\mathcal{S}\times\mathcal{A}$. Here, (i) follows by construction (42b), (ii) invokes the monotonicity property (33) (so that $Q_\tau^{(t)} \ge Q_\tau^{(t-1)}$), and (iii) follows by repeating the arguments (i) and (ii) recursively. Combining the preceding two bounds with the expression (A.18), we conclude that

$$0 \le Q_\tau^\star(s,a) - Q_\tau^{(t+1)}(s,a) \le \gamma\left\|Q_\tau^\star - \tau\log\xi^{(t+1)}\right\|_\infty + \gamma\alpha^{t+1}\left\|Q_\tau^{(0)} - \tau\log\xi^{(0)}\right\|_\infty \tag{A.19}$$

for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, thus concluding the proof.

C.4. Proof of Lemma 4

Recall that, in this scenario, the policies are updated using inexact policy evaluation via (26), namely,

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A},\qquad \pi^{(t+1)}(a|s)=\frac{\big(\pi^{(t)}(a|s)\big)^{1-\frac{\eta\tau}{1-\gamma}}\exp\big(\frac{\eta}{1-\gamma}\widehat{Q}_\tau^{(t)}(s,a)\big)}{\widehat{Z}^{(t)}(s)}, \tag{A.20}$$

where $\widehat{Z}^{(t)}(s):=\sum_{a'}\pi^{(t)}(a'|s)^{1-\frac{\eta\tau}{1-\gamma}}\exp\big(\frac{\eta}{1-\gamma}\widehat{Q}_\tau^{(t)}(s,a')\big)$. To facilitate analysis, we further introduce another auxiliary policy sequence $\{\breve{\pi}^{(t)}\}$, which corresponds to the policy update as if we had access to exact soft Q-function of $\pi^{(t)}$ in the t-th iteration; this is defined as

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A},\qquad \breve{\pi}^{(t+1)}(a|s)=\frac{\big(\pi^{(t)}(a|s)\big)^{1-\frac{\eta\tau}{1-\gamma}}\exp\big(\frac{\eta}{1-\gamma}Q_\tau^{(t)}(s,a)\big)}{Z^{(t)}(s)}, \tag{A.21}$$

where we abuse the notation by letting $Z^{(t)}(s):=\sum_{a'}\pi^{(t)}(a'|s)^{1-\frac{\eta\tau}{1-\gamma}}\exp\big(\frac{\eta}{1-\gamma}Q_\tau^{(t)}(s,a')\big)$. It is worth emphasizing that $\breve{\pi}^{(t+1)}$ is produced on the basis of $\pi^{(t)}$ as opposed to $\breve{\pi}^{(t)}$; it should be viewed as a one-step perfect update from a given policy $\pi^{(t)}$.

We first make note of the following fact: for any step size $0<\eta\le(1-\gamma)/\tau$, it follows from (A.6) — together with the construction (A.20) and (A.21) — that

$$\begin{aligned}
&\big\|\log\pi^{(t+1)}-\log\breve{\pi}^{(t+1)}\big\|_\infty\\
&\le 2\left\|\log\left(\pi^{(t)}(a|s)^{1-\eta\tau/(1-\gamma)}\exp\Big(\frac{\eta}{1-\gamma}\widehat{Q}_\tau^{(t)}(s,a)\Big)\right)-\log\left(\pi^{(t)}(a|s)^{1-\eta\tau/(1-\gamma)}\exp\Big(\frac{\eta}{1-\gamma}Q_\tau^{(t)}(s,a)\Big)\right)\right\|_\infty\\
&=\frac{2\eta}{1-\gamma}\big\|\widehat{Q}_\tau^{(t)}-Q_\tau^{(t)}\big\|_\infty.
\end{aligned}\tag{A.22}$$

Next, let us recall the inequality (A.10) in the proof of Lemma 1 under exact policy evaluation $\breve{\pi}^{(t+1)}(\cdot|s)$; when applied to the current setting, it essentially indicates that

$$\begin{aligned}
V_\tau^{(t)}(s_0)&=\mathop{\mathbb{E}}_{a_0\sim\breve{\pi}^{(t+1)}(\cdot|s_0)}\left[\frac{1-\gamma}{\eta}\log Z^{(t)}(s_0)\right]-\frac{1-\gamma}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\big\|\,\breve{\pi}^{(t+1)}(\cdot|s_0)\right)\\
&=\mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[\frac{1-\gamma}{\eta}\log Z^{(t)}(s_0)\right]-\frac{1-\gamma}{\eta}\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\big\|\,\breve{\pi}^{(t+1)}(\cdot|s_0)\right),
\end{aligned}\tag{A.23}$$

where the last step follows since the quantity $Z^{(t)}(s)$ does not depend on a at all. In order to control the first term of (A.23), we invoke the definition of $\breve{\pi}^{(t+1)}(\cdot|s)$ to show that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[\frac{1-\gamma}{\eta}\log Z^{(t)}(s_0)\right]\\
&\overset{\text{(i)}}{=}\mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[-\tau\log\breve{\pi}^{(t+1)}(a_0|s_0)+Q_\tau^{(t)}(s_0,a_0)+\left(\tau-\frac{1-\gamma}{\eta}\right)\big(\log\breve{\pi}^{(t+1)}(a_0|s_0)-\log\pi^{(t)}(a_0|s_0)\big)\right]\\
&=\mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\big[-\tau\log\pi^{(t+1)}(a_0|s_0)+Q_\tau^{(t)}(s_0,a_0)\big]+\left(\tau-\frac{1-\gamma}{\eta}\right)\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0)\,\big\|\,\pi^{(t)}(\cdot|s_0)\right)
\end{aligned}$$

$$
\begin{aligned}
&\quad - \frac{1-\gamma}{\eta} \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)} \left[\log \breve{\pi}^{(t+1)}(a_0|s_0) - \log \pi^{(t)}(a_0|s_0)\right] \\
&\le \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)} \left[-\tau \log \pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0)\right] + \left(\tau - \frac{1-\gamma}{\eta}\right) \mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0) \,\big\|\, \pi^{(t)}(\cdot|s_0)\right) \\
&\qquad + 2\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty, \qquad \text{(A.24)}
\end{aligned}
$$

where the final step results from (A.22). Putting the above bound together with (A.23) guarantees that

$$
\begin{aligned}
V_\tau^{(t)}(s_0) &\le \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)} \left[-\tau \log \pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0)\right] - \frac{1-\gamma}{\eta} \mathsf{KL}\left(\pi^{(t)}(\cdot|s_0) \,\big\|\, \breve{\pi}^{(t+1)}(\cdot|s_0)\right) \\
&\qquad - \left(\frac{1-\gamma}{\eta} - \tau\right) \mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0) \,\big\|\, \pi^{(t)}(\cdot|s_0)\right) + 2\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty \\
&\le \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)} \left[-\tau \log \pi^{(t+1)}(a_0|s_0) + Q_\tau^{(t)}(s_0,a_0)\right] + 2\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty \\
&\le \mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)} \left[-\tau \log \pi^{(t+1)}(a_0|s_0) + r(s_0,a_0) + \gamma \mathop{\mathbb{E}}_{s_1\sim P(\cdot|s_0,a_0)} \left[V_\tau^{(t)}(s_1)\right]\right] + 2\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty.
\end{aligned}
$$

where the last identity makes use of the relation $Q_\tau^{(t)}(s_0,a_0) = r(s_0,a_0) + \gamma \mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}\left[V_\tau^{(t)}(s_1)\right]$. Invoking the above inequality recursively as in the expression (A.12) (see Lemma 1), we can expand it to establish

$$
V_\tau^{(t)}(s_0) \le V_\tau^{(t+1)}(s_0) + 2\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty \sum_{i=0}^{\infty} \gamma^i = V_\tau^{(t+1)}(s_0) + \frac{2}{1-\gamma}\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty.
$$

C.5. Proof of Lemma 5

First of all, we follow the definition (8) of the entropy-regularized value function to deduce that

$$
\begin{aligned}
V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) &= \mathop{\mathbb{E}}_{\substack{s_0\sim\rho, a_i\sim\pi_\tau^\star(\cdot|s_i),\\ s_{i+1}\sim P(\cdot|s_i,a_i),\forall i\ge 0}} \left[\sum_{i=0}^{\infty} \gamma^i \left(r(s_i,a_i) - \tau \log \pi_\tau^\star(a_i|s_i)\right)\right] - V_\tau^{(t)}(\rho) \\
&= \mathop{\mathbb{E}}_{\substack{s_0\sim\rho, a_i\sim\pi_\tau^\star(\cdot|s_i),\\ s_{i+1}\sim P(\cdot|s_i,a_i),\forall i\ge 0}} \left[\sum_{i=0}^{\infty} \gamma^i \left(r(s_i,a_i) - \tau \log \pi_\tau^\star(a_i|s_i) + V_\tau^{(t)}(s_i) - V_\tau^{(t)}(s_i)\right)\right] - V_\tau^{(t)}(\rho) \\
&= \mathop{\mathbb{E}}_{\substack{s_0\sim\rho, a_i\sim\pi_\tau^\star(\cdot|s_i),\\ s_{i+1}\sim P(\cdot|s_i,a_i),\forall i\ge 0}} \left[V_\tau^{(t)}(s_0) + \sum_{i=0}^{\infty} \gamma^i \left(r(s_i,a_i) - \tau \log \pi_\tau^\star(a_i|s_i) + \gamma V_\tau^{(t)}(s_{i+1}) - V_\tau^{(t)}(s_i)\right)\right] - V_\tau^{(t)}(\rho) \\
&\overset{\text{(i)}}{=} \mathop{\mathbb{E}}_{\substack{s_0\sim\rho, a_i\sim\pi_\tau^\star(\cdot|s_i),\\ s_{i+1}\sim P(\cdot|s_i,a_i),\forall i\ge 0}} \left[\sum_{i=0}^{\infty} \gamma^i \left(r(s_i,a_i) - \tau \log \pi_\tau^\star(a_i|s_i) + \gamma V_\tau^{(t)}(s_{i+1}) - V_\tau^{(t)}(s_i)\right)\right] \\
&\overset{\text{(ii)}}{=} \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_\rho^{\pi_\tau^\star}} \left[\sum_a \pi_\tau^\star(a|s) \left(r(s,a) - \tau \log \pi_\tau^\star(a|s) + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)} \left[V_\tau^{(t)}(s')\right] - V_\tau^{(t)}(s)\right)\right] \\
&\overset{\text{(iii)}}{=} \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_\rho^{\pi_\tau^\star}} \left[\sum_a \pi_\tau^\star(a|s) \left(Q_\tau^{(t)}(s,a) - \tau \log \pi_\tau^\star(a|s)\right) - V_\tau^{(t)}(s)\right]. \qquad \text{(A.25)}
\end{aligned}
$$

Here, (i) is due to the definition $V_\tau^{(t)}(\rho) = \mathbb{E}_{s_0\sim\rho}\big[V_\tau^{(t)}(s_0)\big]$, (ii) follows by aggregating terms corresponding to the same state-action pair and the definition of $d_\rho^{\pi_\tau^\star}$ (cf. (5)), whereas (iii) results from the definition (11a) of the regularized Q-function.

To continue, we shall attempt to control each part of (A.25) separately. To begin with, observe that the first part of (A.25) can be bounded by Jensen's inequality, namely,

$$\begin{aligned}\sum_a \pi_\tau^\star(a|s)\left(Q_\tau^{(t)}(s,a) - \tau\log\pi_\tau^\star(a|s)\right) &= \tau\sum_a \pi_\tau^\star(a|s)\log\left(\frac{\exp\left(Q_\tau^{(t)}(s,a)/\tau\right)}{\pi_\tau^\star(a|s)}\right)\\ &\le \tau\log\left(\sum_a \pi_\tau^\star(a|s)\frac{\exp\left(Q_\tau^{(t)}(s,a)/\tau\right)}{\pi_\tau^\star(a|s)}\right)\\ &= \tau\log\left(\sum_a \exp\left(Q_\tau^{(t)}(s,a)/\tau\right)\right). \end{aligned} \tag{A.26}$$

With regards to the second part of (A.25), it is seen from the definition of $\pi^{(t+1)}$ (cf. (17)) that

$$Q_\tau^{(t)}(s,a) = \tau\log\pi_\tau^{(t+1)}(a|s) + \tau\log\left(\sum_a \exp\left(Q_\tau^{(t)}(s,a)/\tau\right)\right), \tag{A.27}$$

thus allowing one to derive

$$\begin{aligned} V_\tau^{(t)}(s) &= \sum_a \pi_\tau^{(t)}(a|s)\left(Q_\tau^{(t)}(s,a) - \tau\log\pi_\tau^{(t)}(a|s)\right)\\ &\overset{\text{(i)}}{=} \tau\sum_a \pi_\tau^{(t)}(a|s)\left\{\log\pi_\tau^{(t+1)}(a|s) + \log\left(\sum_a \exp\left(Q_\tau^{(t)}(s,a)/\tau\right)\right) - \log\pi_\tau^{(t)}(a|s)\right\}\\ &= \tau\log\left(\sum_a \exp\left(Q_\tau^{(t)}(s,a)/\tau\right)\right) + \tau\sum_a \pi_\tau^{(t)}(a|s)\left(\log\pi_\tau^{(t+1)}(a|s) - \log\pi_\tau^{(t)}(a|s)\right)\\ &= \tau\log\left(\sum_a \exp\left(Q_\tau^{(t)}(s,a)/\tau\right)\right) - \tau\mathsf{KL}\Big(\pi_\tau^{(t)}(a|s)\,\Big\|\,\pi_\tau^{(t+1)}(\cdot|s)\Big), \end{aligned} \tag{A.28}$$

where (i) relies on the identity (A.27). Substituting the inequalities (A.26) and (A.28) into the expression (A.25), we can demonstrate with a little algebra that

$$V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) \le \frac{1}{\eta}\mathop{\mathbb{E}}_{s\sim d_\rho^{\pi_\tau^\star}}\left[\mathsf{KL}\Big(\pi^{(t)}(\cdot|s)\,\Big\|\,\pi^{(t+1)}(\cdot|s)\Big)\right].$$

C.6. Proof of Lemma 1

The results of this lemma, or some similar versions, have appeared in prior work (e.g. Mei et al. (2020, Lemma 10) and Agarwal et al. (2020, Lemma 5.6)). We include the proof here primarily for the sake of self-completeness.

Proof of Eqn. (A.2a). The policy gradient of the unregularized value function $V^{\pi_\theta}(s_0)$ is well-known as the policy gradient theorem (Sutton et al. 2000). Here, we deal with a slightly different variant – an entropy-regularized value function $V_\tau^{\pi_\theta}(s_0)$ in the expression (2) with the softmax policy parameterization in (7). Invoking the Bellman equation and recognizing that $V_\tau^{\pi_\theta}(s_0)$ can be viewed as an unregularized value function with instantaneous rewards $r(s,a)-\tau\log\pi_\theta(a|s)$ for any (s,a), we obtain

$$\begin{aligned}
\nabla_\theta V_\tau^{\pi_\theta}(s_0) &= \nabla_\theta\left[\sum_{a_0}\pi_\theta(a_0|s_0)\Big(r(s_0,a_0)-\tau\log\pi_\theta(a_0|s_0)+\gamma\mathop{\mathbb{E}}_{s'\sim P(\cdot|s_0,a_0)}\left[V_\tau^{\pi_\theta}(s')\right]\Big)\right]\\
&\overset{\text{(i)}}{=}\nabla_\theta\left[\sum_{a_0}\pi_\theta(a_0|s_0)\Big(Q_\tau^{\pi_\theta}(s_0,a_0)-\tau\log\pi_\theta(a_0|s_0)\Big)\right]\\
&=\sum_{a_0}\big(\nabla_\theta\pi_\theta(a_0|s_0)\big)\Big(Q_\tau^{\pi_\theta}(s_0,a_0)-\tau\log\pi_\theta(a_0|s_0)\Big)+\sum_{a_0}\pi_\theta(a_0|s_0)\nabla_\theta\Big(Q_\tau^{\pi_\theta}(s_0,a_0)-\tau\log\pi_\theta(a_0|s_0)\Big)\\
&\overset{\text{(ii)}}{=}\sum_{a_0}\Big(\pi_\theta(a_0|s_0)\nabla_\theta\log\pi_\theta(a_0|s_0)\Big)\Big(Q_\tau^{\pi_\theta}(s_0,a_0)-\tau\log\pi_\theta(a_0|s_0)\Big)\\
&\qquad+\sum_{a_0}\pi_\theta(a_0|s_0)\nabla_\theta\Big(r(s_0,a_0)+\gamma\sum_{s_1}P(s_1|s_0,a_0)V_\tau^{\pi_\theta}(s_1)-\tau\log\pi_\theta(a_0|s_0)\Big),
\end{aligned}$$

where (i) relies on the definition (11a) of $Q_\tau^{\pi_\theta}$, and (ii) makes use of the identity

$$\nabla_\theta\pi_\theta(a_0|s_0)=\pi_\theta(a_0|s_0)\nabla_\theta\log\pi_\theta(a_0|s_0)$$

as well as the definition (11a) of $Q_\tau^{\pi_\theta}$. Given that

$$\sum_{a_0}\pi_\theta(a_0|s_0)\nabla_\theta\log\pi_\theta(a_0|s_0)=\sum_{a_0}\nabla_\theta\pi_\theta(a_0|s_0)=\nabla_\theta\Big(\sum_{a_0}\pi_\theta(a_0|s_0)\Big)=\nabla_\theta 1=0 \tag{A.29}$$

and that $r(s,a)$ is independent of θ, one can continue the above derivative to reach

$$\begin{aligned}
\nabla_\theta V_\tau^{\pi_\theta}(s_0) &= \sum_{a_0}\Big(\pi_\theta(a_0|s_0)\nabla_\theta\log\pi_\theta(a_0|s_0)\Big)\Big(Q_\tau^{\pi_\theta}(s_0,a_0)-\tau\log\pi_\theta(a_0|s_0)\Big)\\
&\qquad+\gamma\sum_{a_0}\pi_\theta(a_0|s_0)\sum_{s_1}P(s_1|s_0,a_0)\nabla_\theta V_\tau^{\pi_\theta}(s_1)\\
&=\mathop{\mathbb{E}}_{\substack{a_i\sim\pi_\theta(\cdot|s_i),\\ s_{i+1}\sim P(\cdot|s_i,a_i),\forall i\geq 0}}\Big[\big(\nabla_\theta\log\pi_\theta(a_0|s_0)\big)\Big(Q_\tau^{\pi_\theta}(s_0,a_0)-\tau\log\pi_\theta(a_0|s_0)\Big)+\gamma\nabla_\theta V_\tau^{\pi_\theta}(s_1)\Big].
\end{aligned}$$

Repeating the above calculations recursively, we arrive at

$$\begin{aligned}
\nabla_\theta V_\tau^{\pi_\theta}(s_0) &= \mathop{\mathbb{E}}_{\substack{a_i\sim\pi_\theta(\cdot|s_i),\\ s_{i+1}\sim P(\cdot|s_i,a_i),\forall i\geq 0}}\left[\sum_{t=0}^{\infty}\gamma^t\big(\nabla_\theta\log\pi_\theta(a_t|s_t)\big)\Big(Q_\tau^{\pi_\theta}(s_t,a_t)-\tau\log\pi_\theta(a_t|s_t)\Big)\right]\\
&=\frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_{s_0}^{\pi_\theta}}\mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot|s)}\Big[\big(\nabla_\theta\log\pi_\theta(a|s)\big)\Big(Q_\tau^{\pi_\theta}(s,a)-\tau\log\pi_\theta(a|s)\Big)\Big]
\end{aligned}$$

$$= \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_{s_0}^{\pi_\theta}} \mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot|s)} \Big[\big(\nabla_\theta \log \pi_\theta(a|s)\big) \Big(A_\tau^{\pi_\theta}(s,a) + V_\tau^{\pi_\theta}(s) \Big) \Big]$$
$$= \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_{s_0}^{\pi_\theta}} \mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot|s)} \Big[\big(\nabla_\theta \log \pi_\theta(a|s)\big) A_\tau^{\pi_\theta}(s,a) \Big], \tag{A.30}$$

where the second line follows by aggregating the terms corresponding to the same state-action pair, and the third line invokes the definition (A.1) of $A_\tau^{\pi_\theta}$. To see why the last line holds, invoke (A.29) to reach

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \Big[V_\tau^{\pi_\theta}(s) \nabla_\theta \log \pi_\theta(a|s) \Big] = \sum_a V_\tau^{\pi_\theta}(s) \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)$$
$$= V_\tau^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) = 0.$$

Further, it is easily seen that under the softmax parametrization in (7),

$$\frac{\partial \log \pi_\theta(a'|s')}{\partial \theta(s,a)} = \mathbb{1}[s'=s] \big(\mathbb{1}[a'=a] - \pi_\theta(a|s) \big) \tag{A.31}$$

for any $(s,a),(s',a') \in \mathcal{S}\times\mathcal{A}$. Combining with (A.30), it further implies that

$$\frac{\partial V_\tau^{\pi_\theta}(s_0)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \mathbb{E}_{s'\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a'\sim\pi_\theta(\cdot|s')} \left[\frac{\partial \log \pi_\theta(a'|s')}{\partial \theta(s,a)} A_\tau^{\pi_\theta}(s',a') \right]$$
$$= \frac{1}{1-\gamma} \mathbb{E}_{s'\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a'\sim\pi_\theta(\cdot|s')} \left[\Big(\mathbb{1}[s'=s] \big(\mathbb{1}[a'=a] - \pi_\theta(a|s) \big) \Big) A_\tau^{\pi_\theta}(s',a') \right]$$
$$\overset{\text{(i)}}{=} \frac{1}{1-\gamma} \mathbb{E}_{s'\sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a'\sim\pi_\theta(\cdot|s')} \Big[\mathbb{1}\big[(s',a')=(s,a)\big] A_\tau^{\pi_\theta}(s',a') \Big]$$
$$= \frac{1}{1-\gamma} d_{s_0}^{\pi_\theta}(s) \pi_\theta(a|s) A_\tau^{\pi_\theta}(s,a).$$

where (i) follows from $\mathbb{E}_{a'\sim\pi_\theta(\cdot|s')} A_\tau^{\pi_\theta}(s',a') = \sum_{a'} \pi_\theta(a'|s') A_\tau^{\pi_\theta}(s',a') = 0$ due to the definition (A.1). The proof regarding $V_\tau^{\pi_\theta}(\rho)$ can be obtained by averaging the initial state s_0 over the distribution ρ.

Proof of Eqn. (A.2b). In order to establish (A.2b), a crucial observation is that $w_\theta := \big(\mathcal{F}_\rho^\theta\big)^\dagger \nabla_\theta V_\tau^{\pi_\theta}(\rho)$ is exactly the solution to the following least-squares problem

$$\text{minimize}_{w\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \quad \big\| \mathcal{F}_\rho^\theta w - \nabla_\theta V_\tau^{\pi_\theta}(\rho) \big\|_2^2. \tag{A.32}$$

From the definition (14) of the Fisher information matrix, we have

$$\mathcal{F}_\rho^\theta w = \mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \Big[\big(\nabla_\theta \log \pi_\theta(a|s)\big) \big(\nabla_\theta \log \pi_\theta(a|s)\big)^\top w \Big].$$

for any fixed vector $w = [w_{s,a}]_{(s,a)\in\mathcal{S}\times\mathcal{A}}$. As a result, for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ one has

$$\big(\mathcal{F}_\rho^\theta w\big)_{s,a} = \mathbb{E}_{s'\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a'\sim\pi_\theta(\cdot|s')} \left[\frac{\partial \log \pi_\theta(a'|s')}{\partial \theta(s,a)} \left(\sum_{\tilde{s},\tilde{a}} \frac{\partial \log \pi_\theta(a'|s')}{\partial \theta(\tilde{s},\tilde{a})} w_{\tilde{s},\tilde{a}} \right) \right]$$

$$
\begin{aligned}
&\overset{\text{(i)}}{=} \mathbb{E}_{s'\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a'\sim\pi_\theta(\cdot|s')} \left[\mathbb{1}[s'=s] \Big(\mathbb{1}[a'=a] - \pi_\theta(a|s) \Big) \left(\sum_{\tilde{s},\tilde{a}} \mathbb{1}[\tilde{s}=s'] \Big(\mathbb{1}[\tilde{a}=a'] - \pi_\theta(\tilde{a}|\tilde{s}) \Big) w_{\tilde{s},\tilde{a}} \right) \right] \\
&= \mathbb{E}_{s'\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a'\sim\pi_\theta(\cdot|s')} \left[\mathbb{1}[s'=s] \Big(\mathbb{1}[a'=a] - \pi_\theta(a|s) \Big) \left(w_{s',a'} - \sum_{\tilde{a}} \pi_\theta(\tilde{a}|s') w_{s',\tilde{a}} \right) \right] \\
&= d_\rho^{\pi_\theta}(s) \mathbb{E}_{a'\sim\pi_\theta(\cdot|s')} \left[\Big(\mathbb{1}[a'=a] - \pi_\theta(a|s) \Big) \big(w_{s,a'} - c(s) \big) \right] \\
&= d_\rho^{\pi_\theta}(s) \mathbb{E}_{a'\sim\pi_\theta(\cdot|s')} \Big[\mathbb{1}[a'=a] w_{s,a'} - \pi_\theta(a|s) w_{s,a'} - \mathbb{1}[a'=a] c(s) + \pi_\theta(a|s) c(s) \Big] \\
&= d_\rho^{\pi_\theta}(s) \Big[\pi_\theta(a|s) w_{s,a} - \pi_\theta(a|s) c(s) - \pi_\theta(a|s) c(s) + \pi_\theta(a|s) c(s) \Big] \\
&= d_\rho^{\pi_\theta}(s) \pi_\theta(a|s) \big[w_{s,a} - c(s) \big],
\end{aligned}
$$

where (i) makes use of the derivative calculation (A.31), and we define $c(s) := \sum_a \pi_\theta(a|s) w_{s,a}$. Consequently, the objective function of (A.32) can be written as

$$
\begin{aligned}
\left\| \mathcal{F}_\rho^\theta w - \nabla_\theta V_\tau^{\pi_\theta}(\rho) \right\|_2^2 &= \sum_{s,a} \left(d_\rho^{\pi_\theta}(s) \pi_\theta(a|s) \left[w_{s,a} - c(s) \right] - \frac{1}{1-\gamma} d_\rho^{\pi_\theta}(s) \pi_\theta(a|s) A_\tau^{\pi_\theta}(s,a) \right)^2 \\
&= \sum_{s,a} \left(d_\rho^{\pi_\theta}(s) \pi_\theta(a|s) \left(w_{s,a} - c(s) - \frac{1}{1-\gamma} A_\tau^{\pi_\theta}(s,a) \right) \right)^2,
\end{aligned}
$$

which is minimized by choosing $w_{s,a} = \frac{1}{1-\gamma} A_\tau^{\pi_\theta}(s,a) + c(s)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. This concludes the proof.

D. Convergence guarantees for CPI-style policy updates

Employing the SPI update as the improved policy, we arrive at the following CPI-style update

$$
\pi^{(t+1)} = (1-\beta)\pi^{(t)} + \beta \overline{\pi}^{(t+1)}. \tag{A.33a}
$$

Here, $\overline{\pi}^{(t+1)}$ corresponds to a one-step SPI update from $\pi^{(t)}$, namely,

$$
\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \qquad \overline{\pi}^{(t+1)}(a|s) = \frac{1}{\overline{Z}^{(t)}(s)} \exp\left(Q_\tau^{(t)}(s,a)/\tau \right), \tag{A.33b}
$$

where we denote

$$
\overline{Z}^{(t)}(s) = \sum_{a\in\mathcal{A}} \exp\left(Q_\tau^{(t)}(s,a)/\tau \right) \qquad \text{and} \qquad Q_\tau^{(t)} = Q_\tau^{\pi^{(t)}}
$$

as usual. Here, $\beta \in (0,1]$ is a parameter that controls the "conservatism" of the updates. We characterize the convergence rate of this update rule (A.33) in the following theorem.

Theorem 1 (Linear convergence of CPI-style updates) *For any $0 < \beta \le 1$, the update rule (A.33) satisfies*

$$
V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) \le \left\| \frac{\rho}{\mu_\tau^\star} \right\|_\infty \big(1 - \beta(1-\gamma) \big)^t \left(V_\tau^\star(\mu_\tau^\star) - V_\tau^{(0)}(\mu_\tau^\star) \right), \qquad \forall t \ge 0, \tag{A.34}
$$

where $\mu_\tau^\star$ is the stationary distribution defined in (29).

According to Theorem 1, it takes the CPI-style policy update (A.33) at most

$$\frac{1}{\beta(1-\gamma)}\log\left(\left\|\frac{\rho}{\mu_\tau^\star}\right\|_\infty \frac{V_\tau^\star(\mu_\tau^\star)-V_\tau^{(0)}(\mu_\tau^\star)}{\epsilon}\right)$$

iterations to reach $V_\tau^\star(\rho)-V_\tau^{(t)}(\rho)\le\epsilon$. As it turns out, the CPI-style update rule can be analyzed using our framework through the following performance improvement lemma, which is an adaptation of Lemma 1. In what follows, we use $\overline{Q}_\tau^{(t+1)}$ and $\overline{V}_\tau^{(t+1)}$ to abbreviate $Q_\tau^{\overline{\pi}^{(t+1)}}$ and $V_\tau^{\overline{\pi}^{(t+1)}}$, respectively.

Lemma 2 (Performance improvement of CPI-style updates) *Consider the policy update rule* (A.33a) *with any* $\beta\in(0,1]$. *For any distribution* ρ, *one has*

$$V_\tau^{(t+1)}(\rho)-V_\tau^{(t)}(\rho)\ge\frac{\beta\tau}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_\rho^{(t+1)}}\left[\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right)\right].$$

S ee Appendix D.1.

Combining the above result with Lemma 5 and following a similar approach to (60) give

$$\begin{aligned}
V_\tau^\star(\rho)-V_\tau^{(t+1)}(\rho) &= V_\tau^\star(\rho)-V_\tau^{(t)}(\rho)+\left(V_\tau^{(t)}(\rho)-V_\tau^{(t+1)}(\rho)\right)\\
&\overset{\text{(i)}}{\le} V_\tau^\star(\rho)-V_\tau^{(t)}(\rho)-\frac{\beta\tau}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_\rho^{(t+1)}}\left[\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right)\right]\\
&\overset{\text{(ii)}}{\le} V_\tau^\star(\rho)-V_\tau^{(t)}(\rho)-\frac{\beta\tau}{1-\gamma}\left\|\frac{d_\rho^{\pi_\tau^\star}}{d_\rho^{(t+1)}}\right\|_\infty^{-1}\mathop{\mathbb{E}}_{s\sim d_\rho^{\pi_\tau^\star}}\left[\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right)\right]\\
&\overset{\text{(iii)}}{\le} V_\tau^\star(\rho)-V_\tau^{(t)}(\rho)-\beta\left\|\frac{d_\rho^{\pi_\tau^\star}}{d_\rho^{(t+1)}}\right\|_\infty^{-1}\left(V_\tau^\star(\rho)-V_\tau^{(t)}(\rho)\right)\\
&= \left(1-\beta\left\|\frac{d_\rho^{\pi_\tau^\star}}{d_\rho^{(t+1)}}\right\|_\infty^{-1}\right)\left(V_\tau^\star(\rho)-V_\tau^{(t)}(\rho)\right). \qquad \text{(A.35)}
\end{aligned}$$

Here, (i) arises from Lemma 2, (ii) employs the pre-factor $\left\|d_\rho^{\pi_\tau^\star}/d_\rho^{(t+1)}\right\|_\infty^{-1}$ to accommodate the change of distributions, whereas (iii) follows from Lemma 5 and the constraint that $0\le\eta\le\frac{1-\gamma}{\tau}$. By taking ρ to be the stationary distribution $\mu_\tau^\star$ (cf. (29)), one has

$$\begin{aligned}
V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t+1)}(\mu_\tau^\star) &\le \left(1-\beta\left\|\frac{d_{\mu_\tau^\star}^{\pi_\tau^\star}}{d_{\mu_\tau^\star}^{(t+1)}}\right\|_\infty^{-1}\right)\left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t)}(\mu_\tau^\star)\right)\\
&\le \left(1-\beta\left\|\frac{\mu_\tau^\star}{(1-\gamma)\mu_\tau^\star}\right\|_\infty^{-1}\right)\left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t)}(\mu_\tau^\star)\right)\\
&= (1-\beta(1-\gamma))\left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t)}(\mu_\tau^\star)\right),
\end{aligned}$$

where we have used $d_{\mu_\tau^\star}^{\pi_\tau^\star}=\mu_\tau^\star$ (cf. (29)) and $d_{\mu_\tau^\star}^{(t+1)}\ge(1-\gamma)\mu_\tau^\star$ in the second step. This immediately concludes the proof.

D.1. Proof of Lemma 2

First of all, we claim that

$$V_\tau^{(t+1)}(\rho)-V_\tau^{(t)}(\rho)=\frac{\tau}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_\rho^{(t+1)}}\left[\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right)-\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right)\right], \tag{A.36}$$

which we shall establish momentarily. Since the KL divergence $\mathsf{KL}\left(\pi(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right)$ is convex in $\pi(\cdot|s)$ (Cover 1999), the update rule (A.33a) together with Jensen's inequality necessarily implies that

$$\begin{aligned}\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right)&\le\beta\mathsf{KL}\left(\overline{\pi}^{(t+1)}(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right)+(1-\beta)\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right)\\&=(1-\beta)\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right).\end{aligned}$$

Substituting the above inequality into (A.36) allows us to conclude that

$$V_\tau^{(t+1)}(\rho)-V_\tau^{(t)}(\rho)\ge\frac{\beta\tau}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_\rho^{(t+1)}}\left[\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right)\right].$$

The rest of this proof is then dedicated to establishing the claim (A.36), which is similar to the proof of Lemma 1. To begin with, we express $V_\tau^{(t)}(s_0)$ as follows

$$\begin{aligned}V_\tau^{(t)}(s_0)&=\mathop{\mathbb{E}}_{a_0\sim\pi^{(t)}(\cdot|s_0)}\left[-\tau\log\pi^{(t)}(a_0|s_0)+Q_\tau^{(t)}(s_0,a_0)\right]\\&=\mathop{\mathbb{E}}_{a_0\sim\pi^{(t)}(\cdot|s_0)}\left[-\tau\log\pi^{(t)}(a_0|s_0)+\tau\log\overline{\pi}^{(t+1)}(a_0|s_0)\right]+\tau\log\overline{Z}^{(t)}(s_0)\\&=\tau\log\overline{Z}^{(t)}(s_0)-\tau\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s_0)\right)\\&=\tau\mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[\log\overline{Z}^{(t)}(s_0)\right]-\tau\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s_0)\right),\end{aligned}$$

where the first line makes use of the definitions (8) and (11a), the second line follows from (A.33), the third line uses the definition of the KL divergence, and the last line follows since $\overline{Z}^{(t)}(s_0)$ does not depend on a. To continue, we subtract and add $\tau\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s_0)\right)$ to obtain

$$\begin{aligned}V_\tau^{(t)}(s_0)&=\mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[\tau\log\overline{Z}^{(t)}(s_0)-\tau\log\pi^{(t+1)}(a_0|s_0)+\tau\log\overline{\pi}^{(t+1)}(a_0|s_0)\right]\\&\qquad+\tau\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s_0)\right)-\tau\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s_0)\right)\\&=\mathop{\mathbb{E}}_{a_0\sim\pi^{(t+1)}(\cdot|s_0)}\left[-\tau\log\pi^{(t+1)}(a_0|s_0)+Q_\tau^{(t)}(s_0,a_0)\right]+\tau\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s_0)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s_0)\right)\\&\qquad-\tau\mathsf{KL}\left(\pi^{(t)}(\cdot|s_0)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s_0)\right)\\&=V_\tau^{(t+1)}(s_0)+\frac{\tau}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d_{s_0}^{(t+1)}}\left[\mathsf{KL}\left(\pi^{(t+1)}(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right)-\mathsf{KL}\left(\pi^{(t)}(\cdot|s)\,\|\,\overline{\pi}^{(t+1)}(\cdot|s)\right)\right].\end{aligned}$$

Here, the first step relies on the definition of KL divergence, the second step comes from (A.33), while the last step is obtained by using the relation $Q_\tau^{(t)}(s_0,a_0)=r(s_0,a_0)+\gamma\mathbb{E}_{s_1\sim P(\cdot|s_0,a_0)}\left[V_\tau^{(t)}(s_1)\right]$ and then invoking the above equality recursively as in the expression (A.12) (see Lemma 1). Averaging the equality over the initial state distribution $s_0\sim\rho$ thus establishes the claim (A.36).

E. Proof for approximate entropy-regularized NPG (Theorem 2)

In this section, we complete the proofs of Theorem 2 in Section 4.3, which consists of (i) establishing the linear system in (58) and (ii) extracting the convergence rate from (58).

Step 1: establishing the linear system (58). In what follows, we shall justify the linear system relation by checking each row separately.

(1) **Bounding** $\|Q_\tau^\star - \tau\log\widehat{\xi}^{(t+1)}\|_\infty$**.** From the construction (57b) of $\widehat{\xi}^{(t+1)}$, we have

$$Q_\tau^\star - \tau\log\widehat{\xi}^{(t+1)} = \alpha\big(Q_\tau^\star - \tau\log\widehat{\xi}^{(t)}\big) + (1-\alpha)\big(Q_\tau^\star - Q_\tau^{(t)}\big) + (1-\alpha)\big(Q_\tau^{(t)} - \widehat{Q}_\tau^{(t)}\big).$$

Taken together with the triangle inequality and the assumption $\big\|Q_\tau^{(t)} - \widehat{Q}_\tau^{(t)}\big\|_\infty \le \delta$, this gives

$$\big\|Q_\tau^\star - \tau\log\widehat{\xi}^{(t+1)}\big\|_\infty \le \alpha\big\|Q_\tau^\star - \tau\log\widehat{\xi}^{(t)}\big\|_\infty + (1-\alpha)\big\|Q_\tau^\star - Q_\tau^{(t)}\big\|_\infty + (1-\alpha)\,\delta. \tag{A.37}$$

(2) **Bounding** $-\min_{s,a}\big(Q_\tau^{(t+1)}(s,a) - \tau\log\widehat{\xi}^{(t+1)}(s,a)\big)$**.** Invoking the definition (57b) of $\widehat{\xi}^{(t+1)}$ again implies that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
&-\Big(Q_\tau^{(t+1)}(s,a) - \tau\log\widehat{\xi}^{(t+1)}(s,a)\Big)\\
&\quad= -\Big(Q_\tau^{(t+1)}(s,a) - \tau\Big(\alpha\log\widehat{\xi}^{(t)}(s,a) + (1-\alpha)\widehat{Q}_\tau^{(t)}(s,a)/\tau\Big)\Big)\\
&\quad= -\alpha\Big(Q_\tau^{(t)}(s,a) - \tau\log\widehat{\xi}^{(t)}(s,a)\Big) + (1-\alpha)\Big(\widehat{Q}_\tau^{(t)}(s,a) - Q_\tau^{(t)}(s,a)\Big) + \big(Q_\tau^{(t)}(s,a) - Q_\tau^{(t+1)}(s,a)\big)\\
&\quad\le -\alpha\Big(Q_\tau^{(t)}(s,a) - \tau\log\widehat{\xi}^{(t)}(s,a)\Big) + (1-\alpha)\,\delta + \frac{2\gamma\delta}{1-\gamma},
\end{aligned}$$

where the last inequality follows from $\big\|Q_\tau^{(t)} - \widehat{Q}_\tau^{(t)}\big\|_\infty \le \delta$ and (56). Taking the maximum over $(s,a)\in\mathcal{S}\times\mathcal{A}$ on both sides and using the definition $\alpha = 1 - \frac{\eta\tau}{1-\gamma}$ yield

$$-\min_{s,a}\Big(Q_\tau^{(t+1)}(s,a) - \tau\log\widehat{\xi}^{(t+1)}(s,a)\Big) \le -\alpha\min_{s,a}\Big(Q_\tau^{(t)}(s,a) - \tau\log\widehat{\xi}^{(t)}(s,a)\Big) + (1-\alpha)\,\delta\left(1 + \frac{2\gamma}{\eta\tau}\right). \tag{A.38}$$

(3) **Bounding** $\big\|Q_\tau^\star - Q_\tau^{(t+1)}\big\|_\infty$**.** Following the same arguments as for (A.18), we obtain

$$\begin{aligned}
Q_\tau^\star(s,a) - Q_\tau^{(t+1)}(s,a) &= \gamma\mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[\tau\log\big(\big\|\exp\big(Q_\tau^\star(s',\cdot)/\tau\big)\big\|_1\big) - \tau\log\big(\big\|\widehat{\xi}^{(t+1)}(s',\cdot)\big\|_1\big)\Big]\\
&\qquad - \gamma\mathop{\mathbb{E}}_{\substack{s'\sim P(\cdot|s,a),\\ a'\sim\pi^{(t+1)}(\cdot|s')}}\Big[Q_\tau^{(t+1)}(s',a') - \tau\log\widehat{\xi}^{(t+1)}(s',a')\Big]\\
&\le \gamma\big\|Q_\tau^\star - \tau\log\widehat{\xi}^{(t+1)}\big\|_\infty - \gamma\min_{s,a}\Big(Q_\tau^{(t+1)}(s,a) - \tau\log\widehat{\xi}^{(t+1)}(s,a)\Big),
\end{aligned}$$

where the last line follows from (A.5). By plugging (A.37) and (A.38) into the above inequality, we arrive at the claimed bound regarding this term.

Step 2: deducing convergence guarantees from the linear system (58). We start by pinning down the eigenvalues and eigenvectors of the matrix B. Specifically, the three eigenvalues can be calculated as

$$\lambda_1 = \alpha + \gamma(1-\alpha) = 1-\eta\tau, \qquad \lambda_2 = \alpha \qquad \text{and} \qquad \lambda_3 = 0, \tag{A.39}$$

whose corresponding eigenvectors are given respectively by

$$v_1 = \begin{bmatrix} \gamma \\ 1 \\ 0 \end{bmatrix}, \qquad v_2 = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}, \qquad \text{and} \qquad v_3 = \begin{bmatrix} \alpha \\ \alpha - 1 \\ 0 \end{bmatrix}. \tag{A.40}$$

With some elementary computation, one can show that z_0 and b introduced in (59) can be related to the eigenvectors of B in the following way:

$$\begin{aligned} z_0 &\le \begin{bmatrix} \left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty \\ \left\|Q_\tau^\star - \tau\log\widehat{\xi}^{(0)}\right\|_\infty \\ \left\|Q_\tau^{(0)} - \tau\log\widehat{\xi}^{(0)}\right\|_\infty \end{bmatrix} \\ &= \frac{1}{1-\eta\tau}\left[(1-\alpha)\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + \alpha\left(\left\|Q_\tau^\star - \tau\log\widehat{\xi}^{(0)}\right\|_\infty + \left\|Q_\tau^{(0)} - \tau\log\widehat{\xi}^{(0)}\right\|_\infty\right)\right] v_1 \\ &\qquad + \left\|Q_\tau^{(0)} - \tau\log\widehat{\xi}^{(0)}\right\|_\infty v_2 + c_z v_3 \\ &\le \frac{1}{1-\eta\tau}\left(\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + 2\alpha\tau\left\|\log\pi_\tau^\star - \log\pi^{(0)}\right\|_\infty\right) v_1 + \left\|Q_\tau^{(0)} - \tau\log\widehat{\xi}^{(0)}\right\|_\infty v_2 + c_z v_3, \end{aligned} \tag{A.41}$$

where c_z is some scalar whose value is immaterial since the eigenvalue corresponding to v_3 is $\lambda_3 = 0$, and the last line follows from the same reasoning for (52). Another userful identity is:

$$b = (1-\alpha)\delta\begin{bmatrix} \gamma\left(2+\frac{2\gamma}{\eta\tau}\right) \\ 1 \\ 1+\frac{2\gamma}{\eta\tau} \end{bmatrix} = (1-\alpha)\delta\left[\left(2+\frac{2\gamma}{\eta\tau}\right)v_1 + \left(1+\frac{2\gamma}{\eta\tau}\right)v_2\right]. \tag{A.42}$$

With these preparations in place, we can now invoke the recursion relationship (58) and the non-negativity of B to obtain

$$\begin{aligned} z_{t+1} &\le B^{t+1}z_0 + \sum_{s=0}^{t} B^{t-s}b \\ &\le B^{t+1}\left[\frac{1}{1-\eta\tau}\left(\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + 2\alpha\tau\left\|\log\pi_\tau^\star - \log\pi^{(0)}\right\|_\infty\right) v_1 + \left\|Q_\tau^{(0)} - \tau\log\widehat{\xi}^{(0)}\right\|_\infty v_2 + c_z v_3\right] \\ &\qquad + (1-\alpha)\delta\sum_{s=0}^{t} B^{t-s}\left[\left(2+\frac{2\gamma}{\eta\tau}\right)v_1 + \left(1+\frac{2\gamma}{\eta\tau}\right)v_2\right] \\ &= \left[\lambda_1^t\left(\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + 2\alpha\tau\left\|\log\pi_\tau^\star - \log\pi^{(0)}\right\|_\infty\right) + (1-\alpha)\delta\left(2+\frac{2\gamma}{\eta\tau}\right)\frac{1-\lambda_1^{t+1}}{1-\lambda_1}\right] v_1 \\ &\qquad + \left[\lambda_2^{t+1}\left\|Q_\tau^{(0)} - \tau\log\widehat{\xi}^{(0)}\right\|_\infty + (1-\alpha)\delta\left(1+\frac{2\gamma}{\eta\tau}\right)\frac{1-\lambda_2^{t+1}}{1-\lambda_2}\right] v_2, \end{aligned}$$

where the eigenvalues and eigenvectors of B are given in (A.39) and (A.40), respectively, and the second inequality relies on (A.41) and (A.42). Note that we are only interested in the first two

entries of the vector z_t. Since the first two entries of the eigenvector v_2 are non-positive, we can safely drop the term involving v_2 in the above inequality to obtain

$$\begin{aligned}
&\begin{bmatrix} \left\|Q_\tau^\star - Q_\tau^{(t+1)}\right\|_\infty \\ \left\|Q_\tau^\star - \tau\log\widehat{\xi}^{(t+1)}\right\|_\infty \end{bmatrix} \\
&\le \left\{ \lambda_1^t \left(\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + 2\alpha\tau \left\|\log\pi_\tau^\star - \log\pi^{(0)}\right\|_\infty \right) + (1-\alpha)\delta \left(2 + \frac{2\gamma}{\eta\tau}\right) \frac{1-\lambda_1^{t+1}}{1-\lambda_1} \right\} \begin{bmatrix}\gamma \\ 1\end{bmatrix} \\
&\le \left\{ (1-\eta\tau)^t \left(\left\|Q_\tau^\star - Q_\tau^{(0)}\right\|_\infty + 2\left(1 - \frac{\eta\tau}{1-\gamma}\right)\tau \left\|\log\pi_\tau^\star - \log\pi^{(0)}\right\|_\infty \right) + \frac{2\delta}{1-\gamma}\left(1 + \frac{\gamma}{\eta\tau}\right) \right\} \begin{bmatrix}\gamma \\ 1\end{bmatrix}.
\end{aligned} \tag{A.43}$$

When it comes to the log policies, we recall again the fact that $\pi^{(t)}$ is related to $\widehat{\xi}^{(t)}$ as

$$\forall s \in \mathcal{S}: \qquad \pi^{(t)}(\cdot|s) = \frac{1}{\left\|\widehat{\xi}^{(t)}(s,\cdot)\right\|_1} \widehat{\xi}^{(t)}(s,\cdot). \tag{A.44}$$

Invoking the elementary property (A.6), we reach

$$\left\|\log\pi_\tau^\star - \log\pi^{(t+1)}\right\|_\infty \le 2\left\|Q_\tau^\star/\tau - \log\widehat{\xi}^{(t+1)}\right\|_\infty.$$

This together with the bound on $\left\|Q_\tau^\star - \tau\log\widehat{\xi}^{(t+1)}\right\|_\infty$ in (A.43) establishes our claim for $\left\|\log\pi_\tau^\star - \log\pi^{(t+1)}\right\|_\infty$.

F. Proof for local quadratic convergence (Theorem 3)

Assuming that the policy $\pi^{(t)}$ obeys Condition (30), we can control the difference of the corresponding discounted state visitation probabilities in terms of the sub-optimality gap w.r.t. the log policy. This is stated in the following lemma, whose proof is deferred to Section F.1.

Lemma 3 *Consider any policy π satisfying $\|\log\pi - \log\pi_\tau^\star\|_\infty \le 1$. It follows that*

$$\left\|1 - \frac{d_\rho^{\pi_\tau^\star}}{d_\rho^\pi}\right\|_\infty \le 2\left(\frac{1}{1-\gamma}\left\|\frac{d_\rho^{\pi_\tau^\star}}{\rho}\right\|_\infty - 1\right)\left\|\log\pi - \log\pi_\tau^\star\right\|_\infty.$$

In particular, by taking $\rho = \mu_\tau^\star$ one has

$$\left\|1 - \frac{\mu_\tau^\star}{d_{\mu_\tau^\star}^\pi}\right\|_\infty = \left\|1 - \frac{d_{\mu_\tau^\star}^{\pi_\tau^\star}}{d_{\mu_\tau^\star}^\pi}\right\|_\infty \le \frac{2\gamma}{1-\gamma}\left\|\log\pi - \log\pi_\tau^\star\right\|_\infty.$$

First, by virtue of the SPI update rule (17) and the inequality (A.6), it is guaranteed that

$$\left\|\log\pi_\tau^\star - \log\pi^{(t+1)}\right\|_\infty \le \frac{2}{\tau}\left\|Q_\tau^\star - Q_\tau^{(t)}\right\|_\infty \le \frac{2\gamma}{\tau}\left\|V_\tau^\star - V_\tau^{(t)}\right\|_\infty \le \frac{2\gamma}{\tau}\left\|\frac{1}{\mu_\tau^\star}\right\|_\infty \left(V_\tau^\star(\mu_\tau^\star) - V_\tau^{(t)}(\mu_\tau^\star)\right), \tag{A.45}$$

where the last inequality comes from a change of distributions argument. Armed with Lemma 3 and the inequality (A.45), we arrive at

$$\left\|1-\frac{d_{\mu_\tau^\star}^{\pi_\tau^\star}}{d_{\mu_\tau^\star}^{(t+1)}}\right\|_\infty \le \frac{2\gamma}{1-\gamma}\left\|\log\pi^{(t+1)}-\log\pi_\tau^\star\right\|_\infty \le \frac{4\gamma^2}{(1-\gamma)\tau}\left\|\frac{1}{\mu_\tau^\star}\right\|_\infty \left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t)}(\mu_\tau^\star)\right). \tag{A.46}$$

Substitution into (60) gives

$$\begin{aligned}
V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t+1)}(\mu_\tau^\star) &\le \left(1-\left\|\frac{d_{\mu_\tau^\star}^{\pi_\tau^\star}}{d_{\mu_\tau^\star}^{(t+1)}}\right\|_\infty^{-1}\right)\left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t)}(\mu_\tau^\star)\right)\\
&= \left(1-\left\|1+\frac{d_{\mu_\tau^\star}^{\pi_\tau^\star}-d_{\mu_\tau^\star}^{(t+1)}}{d_{\mu_\tau^\star}^{(t+1)}}\right\|_\infty^{-1}\right)\left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t)}(\mu_\tau^\star)\right)\\
&\le \left(1-\frac{1}{1+\frac{4\gamma^2}{(1-\gamma)\tau}\left\|\frac{1}{\mu_\tau^\star}\right\|_\infty\left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t)}(\mu_\tau^\star)\right)}\right)\left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t)}(\mu_\tau^\star)\right)\\
&= \frac{\frac{4\gamma^2}{(1-\gamma)\tau}\left\|\frac{1}{\mu_\tau^\star}\right\|_\infty\left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t)}(\mu_\tau^\star)\right)^2}{1+\frac{4\gamma^2}{(1-\gamma)\tau}\left\|\frac{1}{\mu_\tau^\star}\right\|_\infty\left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t)}(\mu_\tau^\star)\right)} \le \frac{4\gamma^2}{(1-\gamma)\tau}\left\|\frac{1}{\mu_\tau^\star}\right\|_\infty\left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t)}(\mu_\tau^\star)\right)^2,
\end{aligned}$$

where the second inequality makes use of the bound (A.46). This in turn reveals that

$$\frac{4\gamma^2}{(1-\gamma)\tau}\left\|\frac{1}{\mu_\tau^\star}\right\|_\infty\left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t+1)}(\mu_\tau^\star)\right) \le \left(\frac{4\gamma^2}{(1-\gamma)\tau}\left\|\frac{1}{\mu_\tau^\star}\right\|_\infty\left(V_\tau^\star(\mu_\tau^\star)-V_\tau^{(t)}(\mu_\tau^\star)\right)\right)^2,$$

which leads to our claimed result by a standard change of distributions.

F.1. Proof of Lemma 3

For any policy π, denote by $P_\pi\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ the state transition matrix induced by π as follows

$$\forall s,s'\in\mathcal{S}: \qquad [P_\pi]_{s,s'} := \mathbb{E}_{a\sim\pi(\cdot|s)}\big[P(s'|s,a)\big]. \tag{A.47}$$

For any policy π satisfying $\|\log\pi-\log\pi_\tau^\star\|_\infty\le 1$, we develop an upper bound on $\left|\left[P_\pi-P_{\pi_\tau^\star}\right]_{s,s'}\right|$ as follows

$$\begin{aligned}
\left|\left[P_\pi-P_{\pi_\tau^\star}\right]_{s,s'}\right| = \left|\sum_a P(s'|s,a)\big(\pi(a|s)-\pi_\tau^\star(a|s)\big)\right| &\le \sum_a P(s'|s,a)\pi_\tau^\star(a|s)\left|\frac{\pi(a|s)}{\pi_\tau^\star(a|s)}-1\right|\\
&\overset{\text{(i)}}{\le} (e-1)\sum_a P(s'|s,a)\pi_\tau^\star(a|s)\left|\log\pi(a|s)-\log\pi_\tau^\star(a|s)\right|\\
&\le \|\log\pi-\log\pi_\tau^\star\|_\infty\,(e-1)\sum_a P(s'|s,a)\pi_\tau^\star(a|s)\\
&\le 2\left[P_{\pi_\tau^\star}\right]_{s,s'}\|\log\pi-\log\pi_\tau^\star\|_\infty,
\end{aligned}$$

where (i) uses the assumption $\|\log\pi^\star-\log\pi\|_\infty\le 1$ together with the elementary inequality $|x|\le (e-1)\,|\log(1+x)|$ when $-1<x\le e-1$. With the preceding bound in mind, we can demonstrate that

$$\left|\left(d_\rho^{\pi_\tau^\star}\right)^\top\left(P_\pi-P_{\pi_\tau^\star}\right)\right| \le 2\,\|\log\pi-\log\pi_\tau^\star\|_\infty\left(d_\rho^{\pi_\tau^\star}\right)^\top P_{\pi_\tau^\star}. \tag{A.48}$$

Here and throughout, we overload the notation $|z|$ for any vector $z \in \mathbb{R}^{|\mathcal{S}|}$ to denote $[|z_i|]_{1\le i\le |\mathcal{S}|}$.

In addition, the definitions of d_ρ^π and $d_\rho^{\pi_\tau^\star}$ admit the following matrix-vector representation:

$$\left(d_\rho^\pi\right)^\top = (1-\gamma)\rho^\top \left(I-\gamma P_\pi\right)^{-1}, \tag{A.49}$$

$$\left(d_\rho^{\pi_\tau^\star}\right)^\top = (1-\gamma)\rho^\top \left(I-\gamma P_{\pi^\star}\right)^{-1}, \tag{A.50}$$

thus allowing one to derive

$$\begin{aligned}\left(d_\rho^\pi - d_\rho^{\pi_\tau^\star}\right)^\top &= (1-\gamma)\rho^\top \left(I-\gamma P_{\pi^\star}\right)^{-1}\left[\left(I-\gamma P_{\pi^\star}\right)-\left(I-\gamma P_\pi\right)\right]\left(I-\gamma P_\pi\right)^{-1} \\ &= \gamma\left(d_\rho^{\pi_\tau^\star}\right)^\top \left(P_\pi - P_{\pi_\tau^\star}\right)\left(I-\gamma P_\pi\right)^{-1}.\end{aligned}$$

This together with the non-negativity of the matrix $(I-\gamma P_\pi)^{-1}$ (Li et al. 2020, Lemma 7) enables the following bound

$$\begin{aligned}\left|\left(d_\rho^\pi - d_\rho^{\pi_\tau^\star}\right)^\top\right| &\le \gamma\left|\left(d_\rho^{\pi_\tau^\star}\right)^\top \left(P_\pi - P_{\pi_\tau^\star}\right)\right|\left(I-\gamma P_\pi\right)^{-1} \\ &\le 2\left\|\log\pi - \log\pi_\tau^\star\right\|_\infty \gamma\left(d_\rho^{\pi_\tau^\star}\right)^\top P_{\pi_\tau^\star}\left(I-\gamma P_\pi\right)^{-1},\end{aligned} \tag{A.51}$$

where the last inequality results from (A.48).

Furthermore, we make the observation that

$$\begin{aligned}\gamma\left(d_\rho^{\pi_\tau^\star}\right)^\top P_{\pi_\tau^\star} &= (1-\gamma)\gamma\rho^\top\left(I-\gamma P_{\pi_\tau^\star}\right)^{-1}P_{\pi_\tau^\star} = (1-\gamma)\gamma\rho^\top\left[\sum_{i=0}^{\infty}\left(\gamma P_{\pi_\tau^\star}\right)^i\right]P_{\pi_\tau^\star} \\ &= (1-\gamma)\rho^\top\left[\sum_{i=1}^{\infty}\left(\gamma P_{\pi_\tau^\star}\right)^i\right] \\ &= (1-\gamma)\rho^\top\left[\left(I-\gamma P_{\pi_\tau^\star}\right)^{-1}-I\right] = \left(d_\rho^{\pi_\tau^\star}\right)^\top - (1-\gamma)\rho^\top \le \left(\left\|\frac{d_\rho^{\pi_\tau^\star}}{\rho}\right\|_\infty - (1-\gamma)\right)\rho^\top,\end{aligned}$$

where the last line comes from a change of distributions argument. Combining this bound with (A.51) gives

$$\begin{aligned}\left|\left(d_\rho^\pi - d_\rho^{\pi_\tau^\star}\right)^\top\right| &\le 2\left\|\log\pi - \log\pi_\tau^\star\right\|_\infty\left(\frac{1}{1-\gamma}\left\|\frac{d_\rho^{\pi_\tau^\star}}{\rho}\right\|_\infty - 1\right)(1-\gamma)\rho^\top\left(I-\gamma P_\pi\right)^{-1} \\ &= 2\left\|\log\pi - \log\pi_\tau^\star\right\|_\infty\left(\frac{1}{1-\gamma}\left\|\frac{d_\rho^{\pi_\tau^\star}}{\rho}\right\|_\infty - 1\right)\left(d_\rho^\pi\right)^\top,\end{aligned}$$

where the last line arises from the expression (A.49). As a result, we establish the claimed bound

$$\left\|1-\frac{d_\rho^{\pi_\tau^\star}}{d_\rho^\pi}\right\|_\infty \le 2\left\|\log\pi - \log\pi_\tau^\star\right\|_\infty\left(\frac{1}{1-\gamma}\left\|\frac{d_\rho^{\pi_\tau^\star}}{\rho}\right\|_\infty - 1\right).$$

References

Agarwal, A., Jiang, N., and Kakade, S. M. (2019). Reinforcement learning: Theory and algorithms. Technical report.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2020). Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *arXiv preprint arXiv:2006.03041*.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. *arXiv preprint arXiv:2005.06392*.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. (2017). Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2775–2785.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063.