A Notations in Theoretical Analysis

We introduce notations used in Appendices.

A.1 General Notations

If two functions f and g satisfy $f(x) > (\geq)\ g(x)$ for each x in the domain of f and g, we write $f > (\geq)\ g$. We denote a constant function by its value. For example, if a function in $\mathcal{Q}$ takes $c \in \mathbb{R}$ at each $(s,a) \in \mathcal{S} \times \mathcal{A}$, the function is denoted as c.

We define the following quantities:

$$\omega := \frac{1}{1-\gamma}, \omega_k := \frac{1-\gamma^k}{1-\gamma}, E_k := \sum_{l=0}^{k} \alpha^l \varepsilon_{k-l}, \text{ and } A_k := \sum_{l=0}^{k-1} \alpha^l,$$

where $A_0 := 0$, and error functions ε_l depend on context, which shall be clear.

A.2 Notations Related to Softmax Operators

We call the following policy a Boltzmann policy (given a function $f \in \mathcal{Q}$):

$$b_\beta(a|s;f) := \frac{\exp\left(\beta f(s,a)\right)}{\sum_{b \in \mathcal{A}} \exp\left(\beta f(s,b)\right)},$$

where $\beta \in (0,\infty)$ is the inverse temperature. A Boltzmann-softmax operator $\boldsymbol{b}_\beta$ is a mapping $f \in \mathcal{Q} \mapsto \boldsymbol{b}_\beta f \in \mathcal{V}$ such that

$$\left(\boldsymbol{b}_\beta f\right)(s) := \sum_{a \in \mathcal{A}} b_\beta(a|s;f) f(s,a)$$

for any state $s \in \mathcal{S}$. Note that the Boltzmann-softmax operator is not a linear operator, as it depends on input f. We define $\boldsymbol{m}_\infty$ and $\boldsymbol{b}_\infty$ to be $\boldsymbol{m}$. As we show later, $\boldsymbol{m}_\beta f \leq \boldsymbol{b}_\beta f$ and $\lim_{\beta \to \infty} \boldsymbol{b}_\beta f = \lim_{\beta \to \infty} \boldsymbol{m}_\beta f = \boldsymbol{m} f$ hold.

A.3 Notations Related to Other Operators

n-th power of an operator $\boldsymbol{O}$ is recursively defined by $\boldsymbol{O}^n Q := \boldsymbol{O}^{n-1}\left(\boldsymbol{O}Q\right)$, where $\boldsymbol{O}^0$ is an identity operator $\boldsymbol{I}$. Linear operators can be expressed as matrices. The addition of linear operators, say $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$, is defined as

$$\boldsymbol{O}_1 + \boldsymbol{O}_2 : f \mapsto \boldsymbol{O}_1 f + \boldsymbol{O}_2 f$$

analogously to the definition of the addition of two matrices. The multiplication of a linear operator $\boldsymbol{O}$ with a scalar c is defined as

$$c\boldsymbol{O} : f \mapsto c\left(\boldsymbol{O} f\right).$$

(Recall that the multiplication of a scalar d and a function g means a function that satisfies $\left(dg\right)(x) := dg(x)$ for any x in the domain of g.)

Suppose a policy π. We define an operator

$$\boldsymbol{P}^{\boldsymbol{\pi}} : f \in \mathcal{Q} \mapsto \boldsymbol{P}\boldsymbol{\pi} f \in \mathcal{Q}$$

and a Bellman operator

$$\boldsymbol{T}^{\boldsymbol{\pi}} : f \in \mathcal{Q} \mapsto r + \gamma \boldsymbol{P}\boldsymbol{\pi} f \in \mathcal{Q}.$$

For any policy π, an operator

$$\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\boldsymbol{\pi}}\right)^{-1} : f \in \mathcal{Q} \mapsto \sum_{t=0}^{\infty} \gamma^t \left(\boldsymbol{P}^{\boldsymbol{\pi}}\right)^t f \in \mathcal{Q}$$

is well defined. As the notation implies, $\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\boldsymbol{\pi}}\right)^{-1}\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\boldsymbol{\pi}}\right) f = f$ holds. We note that $\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\boldsymbol{\pi}}\right)^{-1}$ is a monotone operator. In other words, if $f \geq g$, $\left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\boldsymbol{\pi}}\right)^{-1} f \geq \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\boldsymbol{\pi}}\right)^{-1} g$ holds. Furthermore, we note that $Q^\pi = \left(\boldsymbol{I} - \gamma \boldsymbol{P}^{\boldsymbol{\pi}}\right)^{-1} r$.

B Lemmas on the Mellowmax and Boltzmann-Softmax Operators

Here, we prove lemmas on the mellowmax and Boltzmann-softmax operators.

The following lemma shows a relationship between the mellowmax and Boltzmann-softmax operators.

Lemma 5. *For any inverse temperature $\beta \in (0, \infty)$ and function $f \in \mathcal{Q}$,*

$$\frac{1}{\beta} \log |\mathcal{A}| \geq \boldsymbol{b}_\beta f - \boldsymbol{m}_\beta f \geq 0. \tag{20}$$

Proof. Let $H(s)$ denote the entropy of $b_\beta(\cdot|s; f)$, that is,

$$H(s) := -\sum_{a \in \mathcal{A}} b_\beta(a|s; f) \log b_\beta(a|s; f).$$

It can be rewritten as

$$\begin{aligned} H(s) &= -\sum_{a \in \mathcal{A}} \frac{\exp\left(\beta f(s,a)\right)}{Z(s)} \left(\beta f(s,a) - \log Z(s)\right) \\ &= \log Z(s) - \beta \left(\boldsymbol{b}_\beta f\right)(s) \\ &= \beta \left(\boldsymbol{m}_\beta f\right)(s) - \beta \left(\boldsymbol{b}_\beta f\right)(s) + \log |\mathcal{A}|, \end{aligned}$$

where $Z(s) := \sum_{a \in \mathcal{A}} \exp\left(\beta f(s,a)\right)$, and the last line is obtained by using $(1/\beta) \log\left(Z(s)/|\mathcal{A}|\right) = \left(\boldsymbol{m}_\beta f\right)(s)$. Because $0 \leq H(s) \leq \log |\mathcal{A}|$, the claim holds. □

Lemma 7, which is proven by using the following lemma, states that the mellowmax and Boltzmann-softmax operators are close to the max operator.

Lemma 6. *For any inverse temperature $\beta \in (0, \infty)$, state $s \in \mathcal{S}$ and function $f \in \mathcal{Q}$, $\left(\boldsymbol{m}_\beta f\right)(s)$ is non-decreasing in β while $\left(\boldsymbol{m}_\beta f\right)(s) + \left(\log |\mathcal{A}|\right)/\beta$ is non-increasing in β.*

Proof. The former claim holds since

$$\frac{\partial}{\partial \beta} \left(\boldsymbol{m}_\beta f\right)(s) = \frac{1}{\beta} \left(\left(\boldsymbol{b}_\beta f\right)(s) - \left(\boldsymbol{m}_\beta f\right)(s)\right) \geq 0,$$

where the inequality is due to Lemma 5.

On the other hand,

$$\frac{\partial}{\partial \beta} \left(\left(\boldsymbol{m}_\beta f\right)(s) + \frac{1}{\beta} \log |\mathcal{A}| \right) = \frac{1}{\beta} \left(\left(\boldsymbol{b}_\beta f\right)(s) - \left(\boldsymbol{m}_\beta f\right)(s) - \frac{1}{\beta} \log |\mathcal{A}| \right) \leq 0,$$

where the inequality is again due to Lemma 5. □

Lemma 7. *For any inverse temperature $\beta \in (0, \infty)$ and function $f \in \mathcal{Q}$,*

$$\boldsymbol{m} f - \boldsymbol{b}_\beta f \leq \boldsymbol{m} f - \boldsymbol{m}_\beta f \leq \frac{1}{\beta} \log |\mathcal{A}|.$$

Proof. As shown in Lemma 6, $\left(\boldsymbol{m}_\beta f\right)(s) + \left(\log |\mathcal{A}|\right)/\beta$ is non-increasing in β. Therefore, for any $s \in \mathcal{S}$,

$$\left(\boldsymbol{m}_\beta f\right)(s) + \frac{1}{\beta} \log |\mathcal{A}| \geq \lim_{\beta \to \infty} \left(\boldsymbol{m}_\beta f\right)(s) = \left(\boldsymbol{m} f\right)(s),$$

where the last equality is proven in Asadi and Littman (2017). From Lemma 5, $\boldsymbol{m}_\beta f \leq \boldsymbol{b}_\beta f$, and thus, the claim holds. □

C Proof of theorems

In this appendix, we prove Theorem 1, 2, 4 and Proposition 3.

To begin with, we prove several lemmas used throughout the paper. The following lemma not only makes our theoretical analysis simpler, but also shows that the behavior of CVI is determined by a series of functions whose update rule is simpler.

Lemma 8. *Suppose series of policies p_k and functions $\varepsilon_k \in \mathcal{Q}$, where $k = 0, 1, \ldots$. Define $\Phi_k \in \mathcal{Q}$ recursively by*

$$\Phi_{k+1} := \boldsymbol{T}^{\boldsymbol{p_k}} \Phi_k + \alpha \left(\Phi_k - \boldsymbol{p_k} \Phi_k \right) + \varepsilon_k,$$

where $\Phi_0 \in \mathcal{Q}$. For any positive integer K,

$$\Phi_K = A_K \phi_K + \alpha^K \phi_0 - \alpha \boldsymbol{p}_{K-1} \left(A_{K-1} \phi_{K-1} + \alpha^{K-1} \phi_0 \right) \tag{21}$$

holds, where ϕ_k is recursively defined by $\phi_0 := \Phi_0$ and

$$A_{k+1} \phi_{k+1} := A_k \boldsymbol{T}^{\boldsymbol{p_k}} \phi_k + \alpha^k \boldsymbol{T}^{\boldsymbol{p_k}} \phi_0 + E_k. \tag{22}$$

Proof. We prove the claim by induction. For $K = 1$, $\Phi_1 = \boldsymbol{T}^{\boldsymbol{p_0}} \Phi_0 + \alpha \left(\Phi_0 - \boldsymbol{p}_0 \Phi_0 \right) + \varepsilon_0 = A_1 \phi_1 + \alpha \phi_0 - \alpha \boldsymbol{p}_0 \phi_0$. Therefore, the claim holds for $K = 1$.

Next, suppose that up to $K - 1$ ($K > 1$), the claim holds. Then,

$$\begin{aligned}
\boldsymbol{T}^{\boldsymbol{p_{K-1}}} \Phi_{K-1} &= \boldsymbol{T}^{\boldsymbol{p_{K-1}}} \left[A_{K-1} \phi_{K-1} + \alpha^{K-1} \phi_0 - \alpha \boldsymbol{p}_{K-2} \left(A_{K-2} \phi_{K-2} + \alpha^{K-2} \phi_0 \right) \right] \\
&= \left(A_{K-1} + \alpha^{K-1} - \alpha A_{K-2} - \alpha^{K-1} \right) r \\
&\qquad + \gamma A_{K-1} \boldsymbol{P}^{\boldsymbol{p_{K-1}}} \phi_{K-1} + \alpha^{K-1} \gamma \boldsymbol{P}^{\boldsymbol{p_{K-1}}} \phi_0 - \alpha \gamma A_{K-2} \boldsymbol{P}^{\boldsymbol{p_{K-2}}} \phi_{K-2} - \alpha^{K-1} \gamma \boldsymbol{P}^{\boldsymbol{p_{K-2}}} \phi_0 \\
&= A_{K-1} \boldsymbol{T}^{\boldsymbol{p_{K-1}}} \phi_{K-1} + \alpha^{K-1} \boldsymbol{T}^{\boldsymbol{p_{K-1}}} \phi_0 - \alpha A_{K-2} \boldsymbol{T}^{\boldsymbol{p_{K-2}}} \phi_{K-2} - \alpha^{K-1} \boldsymbol{T}^{\boldsymbol{p_{K-2}}} \phi_0 \\
&= A_K \phi_K - \alpha A_{K-1} \phi_{K-1} - E_{K-1} + \alpha E_{K-2} \\
&= A_K \phi_K - \alpha A_{K-1} \phi_{K-1} - \varepsilon_{K-1},
\end{aligned}$$

where the second line follows because $A_{K-1} - \alpha A_{K-2} = 1$. Furthermore,

$$\Phi_{K-1} - \boldsymbol{p}_{K-1} \Phi_{K-1} = A_{K-1} \phi_{K-1} + \alpha^{K-1} \phi_0 - \boldsymbol{p}_{K-1} \left(A_{K-1} \phi_{K-1} + \alpha^{K-1} \phi_0 \right).$$

In the consequence,

$$\Phi_K = \boldsymbol{T}^{\boldsymbol{p_{K-1}}} \Phi_{K-1} + \alpha \left(\Phi_{K-1} - \boldsymbol{p}_{K-1} \Phi_{K-1} \right) + \varepsilon_{K-1} = A_K \phi_K + \alpha^K \phi_0 - \alpha \boldsymbol{p}_{K-1} \left(A_{K-1} \phi_{K-1} + \alpha^{K-1} \phi_0 \right).$$

This concludes the proof. □

The following lemma is frequently used in our theoretical analysis.

Lemma 9. *Suppose series of functions Φ_k and ϕ_k defined in Lemma 8. For any non-negative integer K and a policy ρ satisfying $\boldsymbol{\rho} \geq_{\Phi_K} \boldsymbol{m}_\beta$,*

$$\boldsymbol{\rho} \left(\frac{A_K}{A_{K+1}} \phi_K + \frac{\alpha^K}{A_{K+1}} \phi_0 \right) \geq \boldsymbol{m} \left(\frac{A_K}{A_{K+1}} \phi_K + \frac{\alpha^K}{A_{K+1}} \phi_0 \right) - \frac{\log |\mathcal{A}|}{\beta A_{K+1}}$$

holds.

Proof. From Lemma 8 and the definition of ρ,

$$\begin{aligned}
\boldsymbol{\rho} \Phi_K &= \boldsymbol{\rho} \left(A_K \phi_K + \alpha^K \phi_0 \right) - \alpha \boldsymbol{p}_{K-1} \left(A_{K-1} \phi_{K-1} + \alpha^{K-1} \phi_0 \right) \\
&\geq \boldsymbol{m}_\beta \Phi_K \\
&= \boldsymbol{m}_\beta \left(A_K \phi_K + \alpha^K \phi_0 \right) - \alpha \boldsymbol{p}_{K-1} \left(A_{K-1} \phi_{K-1} + \alpha^{K-1} \phi_0 \right),
\end{aligned}$$

and thus,

$$\boldsymbol{\rho}\left(A_K\phi_K + \alpha^K\phi_0\right) \geq \boldsymbol{m}_\beta\left(A_K\phi_K + \alpha^K\phi_0\right).$$

As a result,

$$\begin{aligned}\boldsymbol{\rho}\left(\frac{A_K}{A_{K+1}}\phi_K + \frac{\alpha^K}{A_{K+1}}\phi_0\right) &\geq \frac{\boldsymbol{m}_\beta\left(A_K\phi_K + \alpha^K\phi_0\right)}{A_{K+1}}\\ &= \boldsymbol{m}_{\beta A_{K+1}}\left(\frac{A_K}{A_{K+1}}\phi_K + \frac{\alpha^K}{A_{K+1}}\phi_0\right)\\ &\geq \boldsymbol{m}\left(\frac{A_K}{A_{K+1}}\phi_K + \frac{\alpha^K}{A_{K+1}}\phi_0\right) - \frac{\log|\mathcal{A}|}{\beta A_{K+1}},\end{aligned}$$

where the last line follows from Lemma 7. □

C.1 Proof of Theorem 1

We are going to prove Theorem 1. For ease of reading, we state settings in the theorem again.

We suppose $\Psi_k \in \mathcal{Q}$ recursively defined by

$$\Psi_{k+1} := \boldsymbol{T}^{\nu_k}\Psi_k + \alpha\left(\Psi_k - \boldsymbol{\nu}_k\Psi_k\right) + \varepsilon_k$$

where $\Psi_0 \in \mathcal{Q}$, and ν_k is a policy such that $\boldsymbol{\nu_k} \geq_{\Psi_k} \boldsymbol{m}_\beta$. Let ρ_k be a policy such that $\boldsymbol{\rho_k} \geq_{\Psi_k} \boldsymbol{\nu_k}$. The initial function Ψ_0 is assumed to be a constant function whose value is 0. E_k defined with ε_k above is used. We let ψ_k denote a function (ϕ_k) obtained by applying Lemma 8 to Ψ_k.

By the decomposition of $Q^* - Q^{\rho_K}$ to $Q^* - q$ and $-\left(Q^{\rho_K} - q\right)$, where q is some function, it is clear that we need an upper bound of $Q^* - q$ and a lower bound of $Q^{\rho_K} - q$ to show a point-wise upper bound of $Q^* - Q^{\rho_K}$, from which l_p-norm performance bounds can be derived. The following two lemmas give us those upper and lower bounds.

Lemma 10. *Suppose series of functions Ψ_k, ψ_k and policies ρ_k explained in the beginning of this subsection. The following upper bound for $Q^* - \psi_{K+1}$ holds for any non-negative integer K:*

$$Q^* - \psi_{K+1} \leq -\frac{1}{A_{K+1}}\sum_{k=0}^{K}\left(\gamma\boldsymbol{P^*}\right)^k E_{K-k} + \frac{\gamma V_{max}}{A_{K+1}}\sum_{k=0}^{K}\gamma^k\alpha^{K-k} + \frac{\gamma\omega_K}{\beta A_{K+1}}\log|\mathcal{A}|. \tag{23}$$

Proof of Lemma 10. We prove the claim by induction.

Note that $A_K + \alpha^K = 1 + \alpha + \cdots + \alpha^K = A_{K+1}$. Therefore, from (22) and Lemma 9,

$$\psi_{K+1} = r + \gamma\boldsymbol{P^{\nu_K}}\left(\frac{A_K}{A_{K+1}}\psi_K + \frac{\alpha^k}{A_{K+1}}\psi_0\right) + \frac{E_K}{A_{K+1}} \geq r + \gamma\boldsymbol{P}^*\left(\frac{A_K}{A_{K+1}}\psi_K + \frac{\alpha^k}{A_{K+1}}\psi_0\right) + \frac{E_K}{A_{K+1}} - \frac{\gamma\log|\mathcal{A}|}{\beta A_{K+1}}.$$

Accordingly, for $K = 0$,

$$Q^* - \psi_1 = \gamma\boldsymbol{P^*}Q^* - \gamma\boldsymbol{P^{\nu_0}}\psi_0 - \frac{E_0}{A_1} \leq \gamma\boldsymbol{P^*}Q^* - \gamma\boldsymbol{P}^*\psi_0 - \frac{E_0}{A_1} \leq \frac{\gamma V_{max}}{A_1} - \frac{E_0}{A_1},$$

where the last inequality is due to $Q^* \leq V_{max}$. Therefore, the claim holds for $K = 0$.

Suppose that the claim holds up to $K - 1$ ($K > 0$). Then, for any positive integer K,

$$\begin{aligned}Q^* - \psi_{K+1} &\leq \frac{\gamma A_K}{A_{K+1}}\boldsymbol{P^*}\left(Q^* - \psi_K\right) + \frac{\gamma\alpha^K}{A_{K+1}}\boldsymbol{P^*}\left(Q^* - \psi_0\right) - \frac{E_K}{A_{K+1}} + \frac{\gamma\log|\mathcal{A}|}{\beta A_{K+1}}\\ &\leq \frac{\gamma A_K}{A_{K+1}}\boldsymbol{P^*}\left(Q^* - \psi_K\right) + \frac{\gamma\alpha^K}{A_{K+1}}V_{max} - \frac{E_K}{A_{K+1}} + \frac{\gamma\log|\mathcal{A}|}{\beta A_{K+1}},\end{aligned}$$

where the inequalities are obtained similarly to the case in which $K = 0$. By the assumption of the induction,

$$\begin{aligned}
Q^* - \psi_{K+1} &\leq \frac{\gamma A_K}{A_{K+1}} \boldsymbol{P^*} \left(Q^* - \psi_K\right) + \frac{\gamma \alpha^K}{A_{K+1}} V_{max} - \frac{E_K}{A_{K+1}} + \frac{\gamma \log |\mathcal{A}|}{\beta A_{K+1}} \\
&\leq -\frac{1}{A_{K+1}} \sum_{k=0}^{K-1} \left(\gamma \boldsymbol{P^*}\right)^{k+1} E_{K-k-1} + \frac{\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K-1} \gamma^{k+1} \alpha^{K-k-1} + \frac{\gamma^2 \omega_K}{\beta A_{K+1}} \log |\mathcal{A}| \\
&\qquad + \frac{\gamma \alpha^K}{A_{K+1}} V_{max} - \frac{E_K}{A_{K+1}} + \frac{\gamma \log |\mathcal{A}|}{\beta A_{K+1}} \\
&\leq -\frac{1}{A_{K+1}} \sum_{k=0}^{K} \left(\gamma \boldsymbol{P^*}\right)^{k} E_{K-k} + \frac{\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^{k} \alpha^{K-k} + \frac{\gamma \omega_K}{\beta A_{K+1}} \log |\mathcal{A}|.
\end{aligned}$$

Therefore, the claim holds. □

Lemma 11. *Suppose series of functions Ψ_k, ψ_k and policies ρ_k explained in the beginning of this subsection. The following lower bound for $Q^{\rho_K} - \psi_{K+1}$ holds for any non-negative integer K:*

$$Q^{\rho_K} - \psi_{K+1} \geq -\frac{1}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \boldsymbol{Q}_{K,K-k} E_{K-k} - \frac{\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} - \frac{\gamma^2 \omega \omega_K}{\beta A_{K+1}} \log |\mathcal{A}|, \tag{24}$$

where

$$\boldsymbol{Q}_{K,K-k} := \begin{cases} \boldsymbol{I} & \textit{for } k = 0 \\ (\boldsymbol{I} - \gamma \boldsymbol{P^{\rho_K}})^{-1} \boldsymbol{P^{\rho_K}} \boldsymbol{P^{\rho_{K-1}}} \dots \boldsymbol{P^{\rho_{K-k+2}}} \boldsymbol{P^{\rho_{K-k+1}}} (\boldsymbol{I} - \gamma \boldsymbol{P^{\rho_{K-k}}}) & \textit{for } 1 \leq k \leq K \end{cases}.$$

Proof of Lemma 11. We first note that for any non-negative integer K,

$$\boldsymbol{\nu_K} \left(\frac{A_K}{A_{K+1}} \psi_K + \frac{\alpha^K}{A_{K+1}} \psi_0 \right) \leq \boldsymbol{\rho_K} \left(\frac{A_K}{A_{K+1}} \psi_K + \frac{\alpha^K}{A_{K+1}} \psi_0 \right). \tag{25}$$

Indeed, from Lemma 8,

$$\begin{aligned}
\boldsymbol{\nu_K} \left(A_K \psi_K + \alpha^K \psi_0 \right) &= \boldsymbol{\nu_K} \left(\Psi_K + \alpha \boldsymbol{\nu}_{K-1} \left(A_{K-1} \phi_{K-1} + \alpha^{K-1} \phi_0 \right) \right) \\
&\leq \boldsymbol{\rho_K} \Psi_K + \alpha \boldsymbol{\nu}_{K-1} \left(A_{K-1} \phi_{K-1} + \alpha^{K-1} \phi_0 \right) \\
&= \boldsymbol{\rho_K} \left(\Psi_K + \alpha \boldsymbol{\nu}_{K-1} \left(A_{K-1} \phi_{K-1} + \alpha^{K-1} \phi_0 \right) \right) \\
&= \boldsymbol{\rho_K} \left(A_K \psi_K + \alpha^K \psi_0 \right).
\end{aligned}$$

As both ν_K and ρ_K are linear operators, the inequality (25) is obtained by dividing both sides by A_{K+1}.

For any non-negative integer K,

$$\begin{aligned}
&(\boldsymbol{I} - \gamma \boldsymbol{P^{\rho_K}}) \left(Q^{\rho_K} - \psi_{K+1} \right) \\
&= \gamma \boldsymbol{P^{\rho_K}} \left(\boldsymbol{T^{\nu_K}} \left(\frac{A_K}{A_{K+1}} \psi_K + \frac{\alpha^K}{A_{K+1}} \psi_0 \right) + \frac{E_K}{A_{K+1}} \right) - \gamma \boldsymbol{P^{\nu_K}} \left(\frac{A_K}{A_{K+1}} \psi_K + \frac{\alpha^K}{A_{K+1}} \psi_0 \right) - \frac{E_K}{A_{K+1}} \\
&\geq \gamma \boldsymbol{P^{\rho_K}} \left(\boldsymbol{T^{\nu_K}} \left(\frac{A_K}{A_{K+1}} \psi_K + \frac{\alpha^K}{A_{K+1}} \psi_0 \right) + \frac{E_K}{A_{K+1}} \right) - \gamma \boldsymbol{P^{\rho_K}} \left(\frac{A_K}{A_{K+1}} \psi_K + \frac{\alpha^K}{A_{K+1}} \psi_0 \right) - \frac{E_K}{A_{K+1}} \\
&= \gamma \boldsymbol{P^{\rho_K}} \left(\boldsymbol{T^{\nu_K}} \left(\frac{A_K}{A_{K+1}} \psi_K + \frac{\alpha^K}{A_{K+1}} \psi_0 \right) - \left(\frac{A_K}{A_{K+1}} \psi_K + \frac{\alpha^K}{A_{K+1}} \psi_0 \right) \right) - \frac{1}{A_{K+1}} (\boldsymbol{I} - \gamma \boldsymbol{P^{\rho_K}}) E_K,
\end{aligned}$$

where the inequality (25) is used. Accordingly, from Lemma 9,

$$\begin{aligned}
&\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\right)\left(Q^{\rho_K}-\psi_{K+1}\right)\\
&\geq \gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\left(\boldsymbol{T}^{\boldsymbol{\nu_K}}\left(\frac{A_K}{A_{K+1}}\psi_K+\frac{\alpha^K}{A_{K+1}}\psi_0\right)-\left(\frac{A_K}{A_{K+1}}\psi_K+\frac{\alpha^K}{A_{K+1}}\psi_0\right)\right)-\frac{1}{A_{K+1}}\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\right)E_K\\
&\geq \gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\left(\boldsymbol{T}^{\boldsymbol{\rho_{K-1}}}\left(\frac{A_K}{A_{K+1}}\psi_K+\frac{\alpha^K}{A_{K+1}}\psi_0\right)-\left(\frac{A_K}{A_{K+1}}\psi_K+\frac{\alpha^K}{A_{K+1}}\psi_0\right)\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad-\frac{1}{A_{K+1}}\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\right)E_K-\frac{\gamma^2\log|\mathcal{A}|}{\beta A_{K+1}}\\
&=\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_{K-1}}}\right)\left(Q^{\rho_{K-1}}-\frac{A_K}{A_{K+1}}\psi_K-\frac{\alpha^K}{A_{K+1}}\psi_0\right)-\frac{1}{A_{K+1}}\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\right)E_K-\frac{\gamma^2\log|\mathcal{A}|}{\beta A_{K+1}}.
\end{aligned}$$

By noting that $A_K+\alpha^K=A_{K+1}$,

$$\begin{aligned}
&\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_{K-1}}}\right)\left(Q^{\rho_{K-1}}-\frac{A_K}{A_{K+1}}\psi_K-\frac{\alpha^K}{A_{K+1}}\psi_0\right)\\
&=\frac{\gamma A_K}{A_{K+1}}\boldsymbol{P}^{\boldsymbol{\rho_K}}\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_{K-1}}}\right)\left(Q^{\rho_{K-1}}-\psi_K\right)+\frac{\gamma\alpha^K}{A_{K+1}}\boldsymbol{P}^{\boldsymbol{\rho_K}}\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_{K-1}}}\right)\left(Q^{\rho_{K-1}}-\psi_0\right)\\
&\geq\frac{\gamma A_K}{A_{K+1}}\boldsymbol{P}^{\boldsymbol{\rho_K}}\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_{K-1}}}\right)\left(Q^{\rho_{K-1}}-\psi_K\right)-\frac{\gamma(1-\gamma)\alpha^K}{A_{K+1}}V_{max},
\end{aligned}$$

where the inequality follows from $\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_{K-1}}}\right)Q^{\rho_{K-1}}=r\geq -r_{max}=-(1-\gamma)V_{max}$. Therefore,

$$\begin{aligned}
&\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\right)\left(Q^{\rho_K}-\psi_{K+1}\right)=\boldsymbol{T}^{\boldsymbol{\rho_K}}\psi_{K+1}-\psi_{K+1}\\
&\geq\frac{\gamma A_K}{A_{K+1}}\boldsymbol{P}^{\boldsymbol{\rho_K}}\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_{K-1}}}\right)\left(Q^{\rho_{K-1}}-\psi_K\right)-\frac{\gamma(1-\gamma)\alpha^K}{A_{K+1}}V_{max}-\frac{1}{A_{K+1}}\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\right)E_K-\frac{\gamma^2\log|\mathcal{A}|}{\beta A_{K+1}}.
\end{aligned}$$

By continuing the same argument, we obtain

$$\begin{aligned}
\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\right)\left(Q^{\rho_K}-\psi_{K+1}\right)&\geq\frac{\gamma^K}{A_{K+1}}\boldsymbol{P}^{\boldsymbol{\rho_K}}\cdots\boldsymbol{P}^{\boldsymbol{\rho_1}}\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_0}}\right)\left(Q^{\rho_0}-\psi_1\right)-\frac{\gamma(1-\gamma)V_{max}}{A_{K+1}}\sum_{k=0}^{K-1}\alpha^{K-k}\gamma^k\\
&\qquad-\frac{1}{A_{K+1}}\sum_{k=0}^{K-1}\gamma^k\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\right)\boldsymbol{Q}_{K,K-k}E_{K-k}-\frac{\gamma^2\omega_K\log|\mathcal{A}|}{\beta A_{K+1}}.
\end{aligned}$$

Since

$$\begin{aligned}
\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_0}}\right)\left(Q^{\rho_0}-\psi_1\right)&=\boldsymbol{T}^{\boldsymbol{\rho_0}}\psi_1-\psi_1\\
&\geq\gamma\boldsymbol{P}^{\boldsymbol{\rho_0}}\left(r+\gamma\boldsymbol{P}^{\boldsymbol{\nu_0}}\psi_0+E_0-\psi_0\right)-E_0\\
&\geq-\gamma(1-\gamma)V_{max}-\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_0}}\right)E_0
\end{aligned}$$

we finally obtain

$$\begin{aligned}
&\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\right)\left(Q^{\rho_K}-\psi_{K+1}\right)\\
&\geq-\frac{\gamma^K}{A_{K+1}}\boldsymbol{P}^{\boldsymbol{\rho_K}}\cdots\boldsymbol{P}^{\boldsymbol{\rho_1}}\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_0}}\right)E_0-\frac{1}{A_{K+1}}\sum_{k=0}^{K-1}\gamma^k\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\right)\boldsymbol{Q}_{K,K-k}E_{K-k}\\
&\qquad\qquad\qquad-\frac{\gamma^{K+1}(1-\gamma)}{A_{K+1}}V_{max}-\frac{\gamma(1-\gamma)V_{max}}{A_{K+1}}\sum_{k=0}^{K-1}\alpha^{K-k}\gamma^k-\frac{\gamma^2\omega_K\log|\mathcal{A}|}{\beta A_{K+1}}\\
&=-\frac{1}{A_{K+1}}\sum_{k=0}^{K}\gamma^k\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\right)\boldsymbol{Q}_{K,K-k}E_{K-k}-\frac{\gamma(1-\gamma)V_{max}}{A_{K+1}}\sum_{k=0}^{K}\alpha^{K-k}\gamma^k-\frac{\gamma^2\omega_K\log|\mathcal{A}|}{\beta A_{K+1}}.
\end{aligned}$$

Recall that $\left(\boldsymbol{I}-\gamma\boldsymbol{P}^{\boldsymbol{\rho_K}}\right)^{-1}$ is monotone and linear. Therefore, by applying it to both sides of the inequality, it is confirmed that the claim holds. □

By combining Lemma 10 and 11, the following proposition is obtained. (Note that the first summation in (26) is from $k = 1$ to K because $\boldsymbol{Q}_{K,K} = (\boldsymbol{P}^*)^0 = \boldsymbol{I}$ for $k = 0$.)

Proposition 12. *Suppose series of functions Ψ_k, ψ_k and policies ρ_k explained in the beginning of this subsection. The following point-wise upper bound for $Q^* - Q^{\rho_K}$ holds for any non-negative integer K:*

$$Q^* - Q^{\rho_K} \leq \frac{1}{A_{K+1}} \sum_{k=1}^{K} \gamma^k \left(\boldsymbol{Q}_{K,K-k} E_{K-k} - (\boldsymbol{P}^*)^k E_{K-k} \right) + \frac{2\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} + \frac{\gamma \omega \omega_K}{\beta A_{K+1}} \log |\mathcal{A}|, \tag{26}$$

and $\sum_{k=1}^{0} Q_k$ means a constant function whose value is 0 for any sequence of functions Q_k.

Now, we prove Theorem 1. We only prove l_∞-norm performance bound. A proof of l_p-norm performance bound is similar to that of (17) given in Appendix C.4. However, we omit it because it is notationally very cluttered.

From Proposition 12 and $|Q^*(s,a) - Q^{\rho_K}(s,a)| = Q^*(s,a) - Q^{\rho_K}(s,a)$,

$$\begin{aligned}
&\|Q^* - Q^{\rho_K}\|_\infty \\
&= \max_{s,a} \left(Q^* - Q^{\pi_K}\right)(s,a) \\
&= \max_{s,a} \left(Q^* - \psi_{K+1} - (Q^{\pi_K} - \psi_{K+1})\right)(s,a) \\
&\leq \max_{s,a} \sum_{k=1}^{K} \frac{\gamma^k}{A_{K+1}} \left(\boldsymbol{Q}_{K,K-k} E_{K-k} - (\boldsymbol{P}^*)^k E_{K-k} \right)(s,a) + \frac{2\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} + \frac{\gamma \omega \omega_K}{\beta A_{K+1}} \log |\mathcal{A}|.
\end{aligned}$$

Because $\|\boldsymbol{Q}_{K,K-k} Q\|_\infty \leq \omega(1+\gamma) \|Q\|_\infty$ for any $Q \in \mathcal{Q}$,

$$\begin{aligned}
\|Q^* - Q^{\rho_K}\|_\infty &\leq \frac{2}{1-\gamma} \sum_{k=1}^{K} \gamma^k \left\| \frac{E_{K-k}}{A_{K+1}} \right\|_\infty + \frac{2\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} + \frac{\gamma \omega \omega_K}{\beta A_{K+1}} \log |\mathcal{A}| \\
&= \frac{2\gamma}{1-\gamma} \sum_{k=0}^{K-1} \gamma^k \left\| \frac{E_{K-k-1}}{A_{K+1}} \right\|_\infty + \frac{2\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} + \frac{\gamma \omega \omega_K}{\beta A_{K+1}} \log |\mathcal{A}|.
\end{aligned}$$

C.2 Proof of Theorem 2

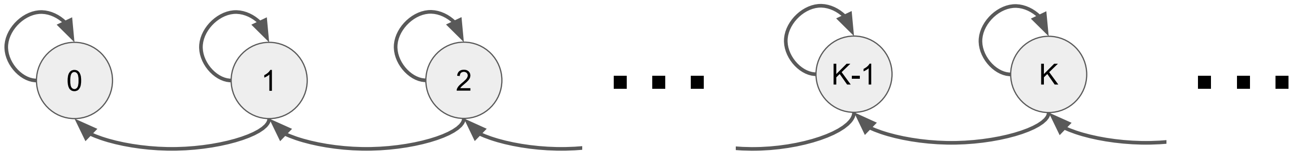

Figure 2: A deterministic environment used to prove the asymptotic tightness of (14) in Theorem 1. This environment is taken from Bertsekas and Tsitsiklis (1996) and Scherrer and Lesner (2012) in which existing performance bounds for VI and policy iteration are proven to be tight. There are two actions: s (stay) and m (move). Except for state 0, staying costs an agent $-r(l,s) = 2\sum_{k=0}^{l-1} \gamma^k \varepsilon$, where $\varepsilon \in (0,\infty)$ is a fixed positive real value, and l is an index of a state. At state 0, no cost is incurred. Therefore, an optimal action is m (move) at all states.

We are going to prove Theorem 2. For ease of reading, we state settings in the theorem again.

Recall that β is assumed to be ∞. Therefore, $\Psi_k \in \mathcal{Q}$ is recursively defined by

$$\Psi_{k+1} := \boldsymbol{T}\Psi_k + \alpha \left(\Psi_k - \boldsymbol{m}\Psi_k\right) + \varepsilon_k$$

where $\Psi_0 \in \mathcal{Q}$. Note that ρ_k is a greedy policy with respect to Ψ_k. The initial function Ψ_0 is assumed to be a constant function whose value is 0. E_k defined with ε_k above is used. We let ψ_k denote a function (ϕ_k) obtained by applying Lemma 8 to Ψ_k.

Since the proof is lengthy, we first provide a sketch of the proof. Then, we provide a full proof.

C.2.1 Proof Sketch

Consider a deterministic environment depicted in Figure 2. Expected immediate reward of staying at state l is given as $r(l,s) = -2\sum_{k=0}^{l-1}\gamma^k\varepsilon$, where $\varepsilon \in (0,\infty)$ is a prescribed positive real value. We assume that

- For any state l and action a, $\Psi_0(l,a) = 0$.
- For any state l and action a, $E_k(l,a) = 0$ except state $l = k+1$ and $l = k+2$ where

$$E_k(k+1,s) = A_{k+1}\varepsilon, \qquad E_k(k+1,m) = -A_{k+1}\varepsilon - \alpha^k\gamma\frac{1-\gamma^k}{1-\gamma}\varepsilon,$$

$$E_k(k+2,s) = 0, \qquad E_k(k+2,m) = A_{k+1}\varepsilon + \alpha^{k+1}\frac{1-\gamma^{k+1}}{1-\gamma}\varepsilon.$$

Under these assumptions, we prove that for any positive integer $K \geq 1$, (i) $\psi_K(K,s) = \psi_K(K,m)$ and (ii) $\psi_K(K+L,s) < \psi_K(K+L,m)$, where $L \in \{1,2,\dots\}$. Thus, from Lemma 8, one of greedy policies with respect to Ψ_K chooses action s (stay) at state K resulting in cumulative rewards of $-2\sum_{t=0}^{\infty}\gamma^t\sum_{k=0}^{K-1}\gamma^k\varepsilon$. We set ρ_K to that greedy policy. As a result,

$$\|Q^* - Q^{\rho_K}\|_\infty = Q^*(K,s) - Q^{\rho_K}(K,s) = \frac{2\gamma(1-\gamma^K)}{(1-\gamma)^2}\varepsilon$$

since $Q^*(K,s) = -2\sum_{k=0}^{K-1}\gamma^k\varepsilon$ is cumulative rewards when s is taken once at state K and m is repeatedly taken afterwards.

On the other hand, it is obvious that either

$$\|E_k\|_\infty = |E_k(k+1,m)| \text{ or } \|E_k\|_\infty = |E_k(k+2,m)|$$

holds. In any case, we have

$$\|E_k\|_\infty = A_{k+1}\varepsilon + O(\alpha^k).$$

Thus, the right hand side of (14) becomes

$$\begin{aligned}
\text{r.h.s.} &= \frac{2\gamma\varepsilon}{1-\gamma}\sum_{k=0}^{K-1}\gamma^{K-k-1}\frac{A_{k+1}}{A_{K+1}} + o(1) \\
&= \frac{2\gamma\varepsilon}{1-\gamma}\sum_{k=0}^{K-1}\gamma^{K-k-1}\left(1 - \alpha^k\frac{A_{K-k+1}}{A_{K+1}}\right) + o(1) \\
&= \frac{2\gamma(1-\gamma^K)}{(1-\gamma)^2}\varepsilon - \frac{2\varepsilon}{(1-\gamma)A_{K+1}}\sum_{k=0}^{K-1}\gamma^k\alpha^{K-k}A_{k+1} + o(1).
\end{aligned}$$

The second term converges to 0. Indeed, when $0 \leq \alpha < 1$,

$$0 \leq \frac{1}{A_{K+1}}\sum_{k=0}^{K-1}\gamma^k\alpha^{K-k}A_{k+1} \leq \alpha^K\sum_{k=0}^{K-1}\left(\frac{\gamma}{\alpha}\right)^k = \frac{\alpha^K - \gamma^K}{\alpha-\gamma}.$$

(When $\alpha = \gamma$, the last term is $K\alpha^K$, which converges to 0.) On the other hand, when $\alpha = 1$,

$$0 \leq \frac{1}{A_{K+1}}\sum_{k=0}^{K-1}\gamma^k\alpha^{K-k}A_{k+1} = \frac{1}{K}\sum_{k=0}^{K-1}\gamma^k(k+1) = \frac{1-\gamma^K}{(1-\gamma)^2K} - \frac{\gamma^K}{1-\gamma},$$

where the second equality is obtained as follows: let S_K denote $\sum_{k=0}^{K-1} \gamma^k (k+1)$. Because

$$\begin{aligned} S_K - \gamma S_K &= \sum_{k=0}^{K-1} \gamma^k (k+1) - \sum_{k=0}^{K-1} \gamma^{k+1} (k+1) \\ &= \sum_{k=0}^{K-1} \gamma^k (k+1) - \sum_{k=1}^{K} \gamma^k k \\ &= \sum_{k=0}^{K-1} \gamma^k - \gamma^K K, \end{aligned}$$

it follows that $S_K = \dfrac{1-\gamma^K}{(1-\gamma)^2} - \dfrac{\gamma^K K}{1-\gamma}$. As a result,

$$\lim_{K\to\infty} \text{r.h.s.} = \frac{2\gamma\varepsilon}{(1-\gamma)^2} = \lim_{K\to\infty} \|Q^* - Q^{\rho_K}\|_\infty.$$

C.2.2 Full Proof

By induction, we prove that for any positive integer $K \geq 1$

$$\psi_K(K,s) = \psi_K(K,m) = -\frac{1-\gamma^K}{1-\gamma}\varepsilon, \tag{27}$$

$$\psi_K(K+1,m) = \frac{A_{K+1}}{A_K}\frac{1-\gamma^K}{1-\gamma}\varepsilon, \tag{28}$$

$$\psi_K(K+L,s) < \psi_K(K+L,m), \tag{29}$$

where $L \in \{1,2,\ldots\}$.

Recall that the update rule of ψ_k is

$$\psi_k = r + \gamma \frac{A_{k-1}}{A_k} \boldsymbol{Pm} \psi_{k-1} + \frac{1}{A_k} E_{k-1},$$

as we assume that $\psi_0 = \Psi_0 = 0$. For $K = 1$, as $\psi_0 = \Psi_0 = 0$,

$$\begin{gathered} \psi_1(1,s) = r(1,s) + E_0(1,s) = -\frac{1-\gamma^1}{1-\gamma}\varepsilon = r(1,m) + E_0(1,m) = \psi_1(1,m) \\ \psi_1(2,m) = r(2,m) + E_0(2,m) = \varepsilon + \alpha\varepsilon = \frac{A_2}{A_1}\frac{1-\gamma^1}{1-\gamma}\varepsilon \\ \psi_1(1+L,s) = r(1+L,s) + E_0(1+L,s) < 0 \leq r(1+L,m) + E_0(1+L,m) = \psi_1(1+L,m). \end{gathered}$$

Therefore, (27), (28) and (29) hold for $K = 1$.

Suppose that (27), (28) and (29) hold up to $K-1$ $(K > 1)$. First, note that

$$\begin{aligned} \psi_K(K,m) &= \gamma \frac{A_{K-1}}{A_K} \max\{\psi_{K-1}(K-1,s), \psi_{K-1}(K-1,m)\} + \frac{1}{A_K} E_{K-1}(K,m) \\ &= -\gamma \frac{A_{K-1}}{A_K} \frac{1-\gamma^{K-1}}{1-\gamma}\varepsilon - \varepsilon - \frac{\alpha^{K-1}}{A_K}\gamma\frac{1-\gamma^{K-1}}{1-\gamma}\varepsilon \\ &= -\varepsilon - \frac{1}{A_K}\left(A_{K-1} + \alpha^{K-1}\right)\gamma\frac{1-\gamma^{K-1}}{1-\gamma}\varepsilon, \\ &= -\frac{1-\gamma^K}{1-\gamma}\varepsilon, \end{aligned}$$

where we used $\max\{\psi_{K-1}(K-1,s), \psi_{K-1}(K-1,m)\} = \psi_{K-1}(K-1,s) = \psi_{K-1}(K-1,m)$ and $A_{K-1}+\alpha^{K-1} = A_K$. Next, note that

$$\begin{aligned}
\psi_K(K,s) &= r(K,s) + \gamma \frac{A_{K-1}}{A_K} \max\{\psi_{K-1}(K,s), \psi_{K-1}(K,m)\} + \frac{1}{A_K} E_{K-1}(K,s) \\
&= -2\frac{1-\gamma^K}{1-\gamma}\varepsilon + \gamma\frac{1-\gamma^{K-1}}{1-\gamma}\varepsilon + \varepsilon \\
&= -\frac{1-\gamma^K}{1-\gamma}\varepsilon,
\end{aligned}$$

where we used $\psi_{K-1}(K,s) < \psi_{K-1}(K,m)$ to obtain $\max\{\psi_{K-1}(K,s), \psi_{K-1}(K,m)\} = \psi_{K-1}(K,m)$. Therefore, (27) holds. Furthermore,

$$\begin{aligned}
\psi_K(K+1,m) &= \gamma \frac{A_{K-1}}{A_K} \max\{\psi_{K-1}(K,s), \psi_{K-1}(K,m)\} + \frac{1}{A_K} E_{K-1}(K+1,m) \\
&= \gamma\frac{1-\gamma^{K-1}}{1-\gamma}\varepsilon + \varepsilon + \frac{\alpha^K}{A_K}\frac{1-\gamma^K}{1-\gamma}\varepsilon \\
&= \left(1+\frac{\alpha^K}{A_K}\right)\frac{1-\gamma^K}{1-\gamma}\varepsilon \\
&= \frac{A_{K+1}}{A_K}\frac{1-\gamma^K}{1-\gamma}\varepsilon,
\end{aligned}$$

where we again used $\psi_{K-1}(K,s) < \psi_{K-1}(K,m)$ to obtain $\max\{\psi_{K-1}(K,s), \psi_{K-1}(K,m)\} = \psi_{K-1}(K,m)$. Thus, (28) holds. Finally, noting that $\psi_{K-1}(K+L-1,s) < \psi_{K-1}(K+L-1,m)$,

$$\begin{aligned}
\psi_K(K+L,m) &= \gamma \frac{A_{K-1}}{A_K} \max\{\psi_{K-1}(K+L-1,s), \psi_{K-1}(K+L-1,m)\} + \frac{1}{A_K} E_{K-1}(K+L,m) \\
&= \gamma \frac{A_{K-1}}{A_K}\psi_{K-1}(K+L-1,m) + \frac{1}{A_K} E_{K-1}(K+L,m) \\
&= \gamma^2 \frac{A_{K-2}}{A_K}\psi_{K-2}(K+L-2,m) + \frac{1}{A_K}\left(E_{K-1}(K+L,m) + \gamma E_{K-2}(K+L-1,m)\right) \\
&\vdots \\
&= \frac{1}{A_K}\left(E_{K-1}(K+L,m) + \gamma E_{K-2}(K+L-1,m) + \cdots + \gamma^{K-1}E_0(L+1,m)\right).
\end{aligned}$$

Because $L \geq 1$, $E_{K-1-i}(K+L-i,m) \geq 0$, and thus, $\psi_K(L,m) \geq 0$. On the other hand,

$$\begin{aligned}
\psi_K(K+L,s) &= r(K+L,s) + \gamma \frac{A_{K-1}}{A_K} \max\{\psi_{K-1}(K+L,s), \psi_{K-1}(K+L,m)\} + \frac{1}{A_K} E_{K-1}(K+L,s) \\
&= r(K+L,s) + \gamma \frac{A_{K-1}}{A_K}\psi_{K-1}(K+L,m) \\
&= r(K+L,s) + \frac{1}{A_K}\left(\gamma E_{K-2}(K+L,m) + \cdots + \gamma^{K-1}E_0(L+2,m)\right).
\end{aligned}$$

Because $L \geq 1$, $E_{K-2-i}(K+L-i,m) = 0$, and thus, $\psi_K(K+L,m) = r(K+L,s) < 0$. Therefore, (29) holds. Given those results,

$$\lim_{K\to\infty} \text{r.h.s.} = \frac{2\gamma\varepsilon}{(1-\gamma)^2} = \lim_{K\to\infty} \|Q^* - Q^{\rho_K}\|_\infty$$

can be shown by following the proof sketch we have provided in Appendix C.2.

C.3 Proof of Proposition 3

We aim at proving Proposition 3. For the ease of reading, we state settings in the proposition again.

We suppose a series of functions $\Psi_k \in \mathcal{Q}$ defined by

$$\Psi_{k+1} := \boldsymbol{T}_\beta \Psi_k + \alpha \left(\Psi_k - \boldsymbol{m_\beta} \Psi_k \right) + \varepsilon_k,$$

where $\Psi_0 \in \mathcal{Q}$. The initial function Ψ_0 is assumed to be a constant function whose value is 0. Furthermore, $\|\varepsilon_l\|_\infty \leq \varepsilon$ is assumed. E_k defined with ε_k above is used. We let ψ_k denote a function (ϕ_k) obtained by applying Lemma 8 to Ψ_k. A policy is given as

$$\pi_{k+1}(a|s) \propto \exp\left(\beta \Psi_{k+1}(s,a)\right) \propto \exp\left(\beta A_{k+1} \psi_{k+1}(s,a)\right).$$

As explained in Asadi and Littman (2017), there exists a policy μ_k such that $\boldsymbol{m_\beta} \Psi_k = \boldsymbol{\mu_k} \Psi_k$. From Lemma 8, it follows that

$$\boldsymbol{m_\beta} \left(A_k \psi_k + \alpha^k \psi_0 \right) = \boldsymbol{\mu_k} \left(A_k \psi_k + \alpha^k \psi_0 \right).$$

Let us start the proof. Since

$$\log \frac{\pi_K(a|s)}{\pi_{K-1}(a|s)} = \beta \left\{ A_K \psi_K(s,a) - A_{K-1} \psi_{K-1}(s,a) - \left[\boldsymbol{m}_\beta \left(A_K \psi_K \right) - \boldsymbol{m}_\beta \left(A_{K-1} \psi_{K-1} \right) \right] \right\},$$

we have (note that the mellowmax is a non-expansion)

$$\left\| \sum_a \pi_K(a|\cdot) \log \frac{\pi_K(a|\cdot)}{\pi_{K-1}(a|\cdot)} \right\|_\infty \leq 2\beta \left\| A_K \psi_K - A_{K-1} \psi_{K-1} \right\|_\infty.$$

By definition, $A_K \psi_K = A_{K-1} \boldsymbol{T}^{\boldsymbol{\mu_{K-1}}} \psi_{K-1} + \alpha^{K-1} r + E_{K-1} = A_K r + \gamma \boldsymbol{P} \boldsymbol{m}_\beta \left(A_{K-1} \psi_{K-1} \right) + E_{K-1}$ as we assumed $\psi_0(s,a) = \Psi_0(s,a) = 0$. Therefore,

$$\begin{aligned} \left\| A_K \psi_K - A_{K-1} \psi_{K-1} \right\|_\infty &= \left\| \alpha^{K-1} r + \gamma \boldsymbol{P} \boldsymbol{m}_\beta \left(A_{K-1} \psi_{K-1} \right) - \gamma \boldsymbol{P} \boldsymbol{m}_\beta \left(A_{K-2} \psi_{K-2} \right) + \varepsilon_{K-1} - (1-\alpha) E_{K-2} \right\|_\infty \\ &\leq \alpha^{K-1} r_{max} + \gamma \left\| A_{K-1} \psi_{K-1} - A_{K-2} \psi_{K-2} \right\|_\infty + 2\varepsilon. \end{aligned}$$

By induction, it is easy to see that

$$\begin{aligned} \left\| A_K \psi_K - A_{K-1} \psi_{K-1} \right\|_\infty &\leq \gamma^{K-1} \left\| A_1 \psi_1 \right\|_\infty + 2(1 + \gamma + \cdots + \gamma^{K-2}) \varepsilon + (\alpha^{K-1} + \alpha^{K-2} \gamma + \cdots + \alpha \gamma^{K-2}) r_{max} \\ &\leq 2 \frac{1-\gamma^K}{1-\gamma} \varepsilon + r_{max} \sum_{k=0}^{K-1} \alpha^k \gamma^{K-k-1}. \end{aligned}$$

As a result, $\left\| \sum_a \pi_K(a|\cdot) \log \frac{\pi_K(a|\cdot)}{\pi_{K-1}(a|\cdot)} \right\|_\infty \leq 4\beta \left(\frac{1-\gamma^K}{1-\gamma} \varepsilon + r_{max} \sum_{k=0}^{K-1} \alpha^k \gamma^{K-k-1} \right)$.

C.4 Proof of Theorem 4

In this appendix, we prove Theorem 4. A basic strategy we take is almost same as the one we used in the proof of Theorem 1. For the ease of reading, we state settings in the theorem again.

We suppose a series of functions $\Psi_k \in \mathcal{Q}$ defined by

$$\Psi_{k+1} := \boldsymbol{T}_\beta \Psi_k + \alpha \left(\Psi_k - \boldsymbol{m_\beta} \Psi_k \right) + \varepsilon_k,$$

where $\Psi_0 \in \mathcal{Q}$. The initial function Ψ_0 is assumed to be a constant function whose value is 0. E_k defined with ε_k above is used. We let ψ_k denote a function (ψ_k) obtained by applying Lemma 8 to Ψ_k. A policy is given as

$$\pi_{k+1}(a|s) \propto \exp\left(\beta \Psi_{k+1}(s,a)\right).$$

As explained in Asadi and Littman (2017), there exists a policy μ_k such that $\boldsymbol{m_\beta} \Psi_k = \boldsymbol{\mu_k} \Psi_k$. From Lemma 8, it follows that

$$\boldsymbol{m_\beta} \left(A_k \psi_k + \alpha^k \psi_0 \right) = \boldsymbol{\mu_k} \left(A_k \psi_k + \alpha^k \psi_0 \right).$$

Dividing both sides by A_{k+1},

$$\boldsymbol{m_{\beta A_{k+1}}} \left(\frac{A_k}{A_{k+1}} \psi_k + \frac{\alpha^k}{A_{k+1}} \psi_0 \right) = \boldsymbol{\mu_k} \left(\frac{A_k}{A_{k+1}} \psi_k + \frac{\alpha^k}{A_{k+1}} \psi_0 \right).$$

We extensively use this equation.

First, we show an upper bound of difference between Q-value functions of two policies.

Lemma 13. *For any pair of policies π and μ, the maximum difference between their Q-value functions is bounded by $\sqrt{2}\gamma\omega V_{max}\delta^{1/2}$, where $\delta = \max_s D_{KL}\left(\pi(\cdot|s)|\mu(\cdot|s)\right)$.*

Proof. We have

$$\begin{aligned} Q^\pi - Q^\mu &= \gamma \boldsymbol{P}^\pi Q^\pi - \gamma \boldsymbol{P}^\mu Q^\mu = \gamma \boldsymbol{P} \left(\boldsymbol{\pi} Q^\pi - \boldsymbol{\mu} Q^\pi \right) + \gamma \boldsymbol{P}^\mu \left(Q^\pi - Q^\mu \right) \\ &= \gamma \left(\boldsymbol{I} - \gamma \boldsymbol{P}^\mu \right)^{-1} \boldsymbol{P} \left(\boldsymbol{\pi} Q^\pi - \boldsymbol{\mu} Q^\pi \right). \end{aligned}$$

Therefore,

$$\begin{aligned} \left\| Q^\pi - Q^\mu \right\|_\infty \le \gamma\omega \left\| \boldsymbol{\pi} Q^\pi - \boldsymbol{\mu} Q^\pi \right\|_\infty &\le \gamma\omega \max_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \left| \left(\pi(a|s) - \mu(a|s)\right) Q^\pi(s,a) \right| \\ &\le \gamma\omega V_{max} \max_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \left| \pi(a|s) - \mu(a|s) \right|, \end{aligned}$$

where the last inequality follows from Hölder's inequality and $\|Q^\pi\|_\infty \le V_{max}$. By Pinsker's inequality, $\max_s \sum_a |\pi(a|s) - \mu(a|s)| \le \sqrt{2}\delta^{1/2}$. In the consequence, $\|Q^\pi - Q^\mu\|_\infty = \sqrt{2}\gamma\omega V_{max}\delta^{1/2}$. □

The following lemma gives us a different upper bound for $Q^{\pi_K} - \psi_{K+1}$.

Lemma 14. *Suppose series of functions Ψ_k, ψ_k and policies π_k explained in the beginning of this subsection. Let δ_k be an upper bound of $\max_s D_{KL}(\pi_k(\cdot|s)|\pi_{k-1}(\cdot|s))$. The following lower bound for $Q^{\pi_K} - \psi_{K+1}$ holds for any non-negative integer K:*

$$Q^{\pi_K} - \psi_{K+1} \ge -\frac{1}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \boldsymbol{P}_{K,K-k+1} E_{K-k} - \frac{\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} - \sqrt{2}\gamma^2\omega V_{max} \sum_{k=0}^{K-1} \gamma^k \frac{A_{K-k}}{A_{K+1}} \delta_{K-k}^{1/2}, \quad (30)$$

where $\sum_{k=1}^{0} Q_k$ means a constant function whose value is 0 for any sequence of functions Q_k, and

$$\boldsymbol{P}_{K,K-k+1} := \begin{cases} \boldsymbol{I} & \textit{for } k = 0 \\ \boldsymbol{P}^{\pi_K} \boldsymbol{P}^{\pi_{K-1}} \cdots \boldsymbol{P}^{\pi_{K-k+2}} \boldsymbol{P}^{\pi_{K-k+1}} & \textit{for } 1 \le k \le K \end{cases}$$

Proof. For any non-negative integer $K \ge 0$,

$$\begin{aligned} Q^{\pi_K} - \psi_{K+1} &= \gamma \boldsymbol{P}^{\pi_K} Q^{\pi_K} - \gamma \boldsymbol{P}^{\mu_K} \left(\frac{A_K}{A_{K+1}} \psi_K + \frac{\alpha^K}{A_{K+1}} \psi_0 \right) - \frac{E_K}{A_{K+1}} \\ &\ge \gamma \frac{A_K}{A_{K+1}} \boldsymbol{P}^{\pi_K} \left(Q^{\pi_{K-1}} - \psi_K \right) - \frac{E_K}{A_{K+1}} - \frac{\gamma V_{max}}{A_{K+1}} \alpha^K + \gamma \frac{A_K}{A_{K+1}} \boldsymbol{P}^{\pi_K} \left(Q^{\pi_K} - Q^{\pi_{K-1}} \right) \\ &\ge \gamma \frac{A_K}{A_{K+1}} \boldsymbol{P}^{\pi_K} \left(Q^{\pi_{K-1}} - \psi_K \right) - \frac{E_K}{A_{K+1}} - \frac{\gamma V_{max}}{A_{K+1}} \alpha^K - \sqrt{2}\gamma^2\omega V_{max} \frac{A_K}{A_{K+1}} \delta_K^{1/2}. \end{aligned}$$

(The first and last term disappear if $K = 0$.) It is clear that the claim holds for $K = 0$. It is not difficult to prove the claim by induction with the aid of the above inequality. □

By combining Lemma 10 and 14, the following proposition is obtained. (Note that the summation in (31) is from $k = 1$ to K because $\boldsymbol{P}_{K,K+1} = \left(\boldsymbol{P}^*\right)^0 = \boldsymbol{I}$ for $k = 0$.)

Proposition 15. *Suppose series of functions Ψ_k, ψ_k and policies π_k explained in the beginning of this subsection. Let δ_k denote an upper bound of $\max_s D_{KL}(\pi_k(\cdot|s)|\pi_{k-1}(\cdot|s))$. The following point-wise upper bound for $Q^* - Q^{\pi_K}$ holds for any non-negative integer K:*

$$Q^* - Q^{\pi_K} \leq \frac{1}{A_{K+1}} \sum_{k=1}^{K} \gamma^k \left(\boldsymbol{P}_{K,K-k+1} E_{K-k} - \left(\boldsymbol{P}^*\right)^k E_{K-k} \right) + \frac{2\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} + \frac{\gamma \omega_K}{\beta A_{K+1}} \log|\mathcal{A}| + \sqrt{2}\gamma^2 \omega V_{max} \sum_{k=0}^{K-1} \gamma^k \frac{A_{K-k}}{A_{K+1}} \delta_{K-k}^{1/2}, \tag{31}$$

where $\boldsymbol{P}_{k,l}$ are defined in Lemma 14, and $\sum_{k=1}^{0} Q_k$ means a constant function whose value is 0 for any sequence of functions Q_k.

Now we prove Theorem 4. We first prove l_∞-norm performance bound. From Proposition 15 and by noting that $|Q^*(s,a) - Q^{\pi_K}(s,a)| = Q^*(s,a) - Q^{\pi_K}(s,a)$,

$$\begin{aligned}
&\|Q^* - Q^{\pi_K}\|_\infty \\
&= \max_{s,a} \left(Q^* - Q^{\pi_K}\right)(s,a) \\
&= \max_{s,a} \left(Q^* - \psi_{K+1} - \left(Q^{\pi_K} - \psi_{K+1}\right)\right)(s,a) \\
&\leq \max_{s,a \in \mathcal{S}\times\mathcal{A}} \sum_{k=1}^{K} \frac{\gamma^k}{A_{K+1}} \left(\boldsymbol{P}_{K,K-k+1} E_{K-k} - \left(\boldsymbol{P}^*\right)^k E_{K-k} \right)(s,a) \\
&\qquad + \frac{2\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} + \frac{\gamma \omega_K}{\beta A_{K+1}} \log|\mathcal{A}| + \sqrt{2}\gamma^2 \omega V_{max} \sum_{k=0}^{K-1} \gamma^k \frac{A_{K-k}}{A_{K+1}} \delta_{K-k}^{1/2}.
\end{aligned}$$

Because $\|\boldsymbol{P}_{K,K-k+1} Q\|_\infty \leq \|Q\|_\infty$ for any $Q \in \mathcal{Q}$,

$$\begin{aligned}
\|Q^* - Q^{\pi_K}\|_\infty &\leq 2\gamma \sum_{k=1}^{K} \gamma^k \left\| \frac{E_{K-k}}{A_{K+1}} \right\|_\infty \\
&\qquad + \frac{2\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} + \frac{\gamma \omega_K}{\beta A_{K+1}} \log|\mathcal{A}| + \sqrt{2}\gamma^2 \omega V_{max} \sum_{k=0}^{K-1} \gamma^k \frac{A_{K-k}}{A_{K+1}} \delta_{K-k}^{1/2}.
\end{aligned}$$

Loosening it by replacing A_{K-k}/A_{K+1} with 1, we conclude the proof for the l_∞-norm performance bound.

Next, we prove l_p-norm performance bound.

$$\begin{aligned}
|Q^* - Q^{\pi_K}| &= Q^* - Q^{\pi_K} \\
&\leq \sum_{k=1}^{K} \frac{\gamma^k}{A_{K+1}} \left(\boldsymbol{P}_{K,K-k+1} E_{K-k} - \left(\boldsymbol{P}^*\right)^k E_{K-k} \right) \\
&\qquad + \frac{2\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} + \frac{\gamma \omega_K}{\beta A_{K+1}} \log|\mathcal{A}| + \sqrt{2}\gamma^2 \omega V_{max} \sum_{k=0}^{K-1} \gamma^k \frac{A_{K-k}}{A_{K+1}} \delta_{K-k}^{1/2} \\
&\leq \sum_{k=1}^{K} \frac{\gamma^k}{A_{K+1}} \left(\boldsymbol{P}_{K,K-k+1} |E_{K-k}| + \left(\boldsymbol{P}^*\right)^k |E_{K-k}| \right) \\
&\qquad + \frac{2\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} + \frac{\gamma \omega_K}{\beta A_{K+1}} \log|\mathcal{A}| + \sqrt{2}\gamma^2 \omega V_{max} \sum_{k=0}^{K-1} \gamma^k \frac{A_{K-k}}{A_{K+1}} \delta_{K-k}^{1/2},
\end{aligned}$$

where $|Q|, Q \in \mathcal{Q}$ is a function such that $|Q|(s,a) = |Q(s,a)|$ for any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$. Because

$f(x) = x^p$ is monotonically increasing in x for any $p \in [0, \infty)$,

$$|Q^* - Q^{\pi_K}|^p \le \left[\sum_{k=1}^{K} \frac{\gamma^k}{A_{K+1}} \left(\boldsymbol{P}_{K,K-k+1} |E_{K-k}| + (\boldsymbol{P^*})^k |E_{K-k}| \right) \right. \\ \left. + \frac{2\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} + \frac{\gamma \omega_K}{\beta A_{K+1}} \log|\mathcal{A}| + \sqrt{2}\gamma^2 \omega V_{max} \sum_{k=0}^{K-1} \gamma^k \frac{A_{K-k}}{A_{K+1}} \delta_{K-k}^{1/2} \right]^p ,$$

where $Q^p, Q \in \mathcal{Q}$ is a function such that $Q^p(s,a) = Q(s,a)^p$ for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. Now, let us introduce variables $\lambda_{1,i}$, $\lambda_{2,i}$, λ_3, λ_4 and $\lambda_{5,i}$. We temporarily do not specify values as they are just introduced to be used with Jensen's inequality as follows:

$$|Q^* - Q^{\pi_K}|^p \le \left(\frac{\Lambda}{A_{K+1}} \right)^p \left[\sum_{k=1}^{K} \left(\frac{\gamma^k \lambda_{1,K-k}}{\Lambda} \boldsymbol{P}_{K,K-k+1} \left| \frac{E_{K-k}}{\lambda_{1,K-k}} \right| + \frac{\gamma^k \lambda_{2,K-k}}{\Lambda} (\boldsymbol{P^*})^k \left| \frac{E_{K-k}}{\lambda_{2,K-k}} \right| \right) \right. \\ \left. + \frac{\lambda_3}{\Lambda} \frac{2\gamma V_{max} \sum_{k=0}^{K} \gamma^k \alpha^{K-k}}{\lambda_3} + \frac{\lambda_4}{\Lambda} \frac{\gamma \omega_K \log|\mathcal{A}|}{\beta \lambda_4} + \sum_{k=0}^{K-1} \frac{\gamma^k \lambda_{5,K-k}}{\Lambda} \frac{\sqrt{2}\gamma^2 \omega \delta_{K-k}^{1/2} A_{K-k} V_{max}}{\lambda_{5,K-k}} \right]^p ,$$

where Λ is a normalization coefficient defined by

$$\Lambda := \sum_{k=1}^{K} \gamma^k \left(\lambda_{1,K-k} + \lambda_{2,K-k} \right) + \lambda_3 + \lambda_4 + \sum_{k=0}^{K-1} \gamma^k \lambda_{5,K-k}.$$

By using Jensens' inequality twice (firstly considering coefficients and secondly considering $\boldsymbol{P}_{i,j}$ as well as $\boldsymbol{P}^*$),

$$|Q^* - Q^{\pi_K}|^p \\ \le \frac{\Lambda^{p-1}}{A_{K+1}^p} \sum_{k=1}^{K} \left(\frac{\gamma^k}{\lambda_{1,K-k}^{p-1}} \boldsymbol{P}_{K,K-k+1} |E_{K-k}|^p + \frac{\gamma^k}{\lambda_{2,K-k}^{p-1}} (\boldsymbol{P^*})^k |E_{K-k}|^p \right) \\ + \frac{\Lambda^{p-1}}{A_{K+1}^p} \left(\frac{\left(2\gamma V_{max} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} \right)^p}{\lambda_3^{p-1}} + \frac{\left(\frac{\gamma \omega_K}{\beta} \log|\mathcal{A}| \right)^p}{\lambda_4^{p-1}} + \sum_{k=0}^{K-1} \gamma^k \frac{\left(\sqrt{2}\gamma^2 \omega \delta_{K-k}^{1/2} A_{K-k} V_{max} \right)^p}{\lambda_{5,K-k}^{p-1}} \right).$$

Now, it is seen that we need l_p-norm bounds of $\sum_{s,a \in \mathcal{S} \times \mathcal{A}} \rho P^{\pi_K} P^{\pi_{K-1}} \cdots P^{\pi_{K-k+2}} P^{\pi_{K-k+1}}(s,a) |E_{K-k}|^p (s,a)$ and $\sum_{s,a \in \mathcal{S} \times \mathcal{A}} \rho P^* \cdots P^*(s,a) |E_{K-k}|^p (s,a)$

Let us focus on $\sum_{s,a \in \mathcal{S} \times \mathcal{A}} \rho P^* \cdots P^*(s,a) |E_{K-k}|^p (s,a)$.

$$\sum_{s,a \in \mathcal{S} \times \mathcal{A}} \rho P^* \cdots P^*(s,a) |E_{K-k}|^p (s,a) = \mathbb{E}_{(S,A) \sim \nu} \left[\frac{\rho P^* \cdots P^*(S,A)}{\nu(S,A)} |E_{K-k}|^p (S,A) \right] \\ \le c(\rho, \nu; \overbrace{\pi^*, \dots, \pi^*}^{k}) \mathbb{E}_{(S,A) \sim \nu} \left[|E_{K-k}|^{2p} (S,A) \right]^{1/2} \\ = c(\rho, \nu; \pi^*, \dots, \pi^*) \|E_{K-k}\|_{\nu,2p}^p ,$$

where Cauchy–Schwarz inequality is used. A bound for $\sum_{s,a \in \mathcal{S} \times \mathcal{A}} \rho P^{\pi_K} \cdots P^{\pi_{K-k+1}}(s,a) |E_{K-k}|^p (s,a)$ can be similarly obtained. Therefore,

$$\mathbb{E}_{(S,A) \sim \rho} |Q^* - Q^{\pi_K}|^p (S,A) \\ \le \frac{\Lambda^{p-1}}{A_{K+1}^p} \sum_{k=1}^{K} \left(\frac{\gamma^k}{\lambda_{1,K-k}^{p-1}} c(\rho, \nu; \pi_K, \dots, \pi_{K-k+1}) + \frac{\gamma^k}{\lambda_{2,K-k}^{p-1}} c(\rho, \nu; \overbrace{\pi^*, \dots, \pi^*}^{k}) \right) \|E_{K-k}\|_{\nu,2p}^p \\ + \frac{\Lambda^{p-1}}{A_{K+1}^p} \left(\frac{\left(2\gamma V_{max} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} \right)^p}{\lambda_3^{p-1}} + \frac{\left(\frac{\gamma \omega_K}{\beta} \log|\mathcal{A}| \right)^p}{\lambda_4^{p-1}} + \sum_{k=0}^{K-1} \gamma^k \frac{\left(\sqrt{2}\gamma^2 \omega \delta_{K-k}^{1/2} A_{K-k} V_{max} \right)^p}{\lambda_{5,K-k}^{p-1}} \right).$$

By setting

$$\begin{aligned}\lambda_{1,K-k} &= c(\rho, \nu; \pi_K, \dots, \pi_{K-k+1})^{1/p} \left\| E_{K-k} \right\|_{\nu,2p}, \\ \lambda_{2,K-k} &= c(\rho, \nu; \overbrace{\pi^*, \dots, \pi^*}^{k})^{1/p} \left\| E_{K-k} \right\|_{\nu,2p}, \\ \lambda_3 &= 2\gamma V_{max} \sum_{k=0}^{K} \gamma^k \alpha^{K-k}, \\ \lambda_4 &= \frac{\gamma \omega_K}{\beta} \log |\mathcal{A}|, \\ \lambda_{5,K-k} &= \sqrt{2}\gamma^2 \omega \delta_{K-k}^{1/2} A_{K-k} V_{max},\end{aligned}$$

we obtain

$$\mathbb{E}_{(S,A)\sim\rho} \left| Q^* - Q^{\pi_K} \right|^p (S, A) \leq \frac{\Lambda^{p-1}}{A_{K+1}^p} \left(\sum_{k=1}^{K} \gamma^k \left(\lambda_{1,K-k} + \lambda_{2,K-k} \right) + \lambda_3 + \lambda_4 + \sum_{k=0}^{K-1} \gamma^k \lambda_{5,K-k} \right) = \frac{\Lambda^p}{A_{K+1}^p}.$$

Accordingly,

$$\begin{aligned}\left\| Q^* - Q^{\pi_K} \right\|_{\rho,p} &\leq \frac{\Lambda}{A_{K+1}} \\ &= \frac{2\gamma}{A_{K+1}} \sum_{k=0}^{K-1} \gamma^k \frac{c(\rho, \nu, 2; \pi_K, \dots, \pi_{K-k})^{1/p} + c(\rho, \nu, 2; \overbrace{\pi^*, \dots, \pi^*}^{k+1})^{1/p}}{2} \left\| E_{K-k-1} \right\|_{\nu,2p} \\ &\quad + \frac{2\gamma V_{max}}{A_{K+1}} \sum_{k=0}^{K} \gamma^k \alpha^{K-k} + \frac{\gamma(1-\gamma^K)}{\beta(1-\gamma)A_{K+1}} \log |\mathcal{A}| + \frac{\sqrt{2}\gamma^2 V_{max}}{1-\gamma} \sum_{k=0}^{K-1} \gamma^k \frac{A_{K-k}}{A_{K+1}} \delta_{K-k}^{1/2}.\end{aligned}$$

Loosening it by replacing A_{K-k}/A_{K+1} with 1 and taking $\sup_{\pi_K, \dots, \pi_0}$, we conclude the proof.