Supplementary Material for "Breaking the Sample Size Barrier in Model-Based Reinforcement Learning"

This supplementary material includes details that support the main text "Breaking the Sample Size Barrier in Model-Based Reinforcement Learning" accepted to Operations Research. In particular, the proofs of all theorems, corollaries and technical lemmas are included.

EC.1. Preliminary facts

We begin by recording a few elementary facts about $\boldsymbol{P}^\pi$ and $\boldsymbol{P}_\pi$ (see definitions in (24)). These are standard results and we omit the proofs for brevity.

LEMMA EC.1. *For any policy π, any probability transition matrix $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ and any $0<\gamma<1$, one has*

(a) $(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1} = \sum_{i=0}^{\infty}(\gamma\boldsymbol{P}_\pi)^i$;

(b) All entries of the matrix $(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}$ are non-negative;

(c) $\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\|_1 \le 1/(1-\gamma)$;

(d) $(1-\gamma)(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\mathbf{1} = \mathbf{1}$;

(e) For any non-negative vectors $\mathbf{0} \le \boldsymbol{r}_1 \le \boldsymbol{r}_2$ of compatible dimension, one has $\mathbf{0} \le (\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\boldsymbol{r}_1 \le (\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\boldsymbol{r}_2$.

The above results continue to hold if $\boldsymbol{P}_\pi$ is replaced by $\boldsymbol{P}^\pi$.

EC.2. Proofs of auxiliary lemmas: infinite-horizon MDPs

EC.2.1. Proofs of Lemma 1 and Lemma 2

Auxiliary notation and preliminaries. Before proceeding, we define several $|\mathcal{S}|$-dimensional auxiliary vectors $\boldsymbol{r}^{(i)}, \boldsymbol{V}^{(i)}, \widehat{\boldsymbol{V}}^{(i)}$ $(1 \le i \le m)$ recursively as follows

$$\begin{array}{lll} \boldsymbol{r}^{(0)} := \boldsymbol{r}_\pi, & \boldsymbol{V}^{(0)} := (\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\boldsymbol{r}^{(0)}, & \widehat{\boldsymbol{V}}^{(0)} := (\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi)^{-1}\boldsymbol{r}^{(0)}, \\ \boldsymbol{r}^{(1)} := \sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\left[\boldsymbol{V}^{(0)}\right]}, & \boldsymbol{V}^{(1)} := (\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\boldsymbol{r}^{(1)}, & \widehat{\boldsymbol{V}}^{(2)} := (\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi)^{-1}\boldsymbol{r}^{(1)}, \\ \quad\vdots & \quad\vdots & \quad\vdots \\ \boldsymbol{r}^{(m)} := \sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\left[\boldsymbol{V}^{(m-1)}\right]}, & \boldsymbol{V}^{(m)} := (\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\boldsymbol{r}^{(m)}, & \widehat{\boldsymbol{V}}^{(m)} := (\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi)^{-1}\boldsymbol{r}^{(m)}, \end{array} \tag{EC.1}$$

where m will be specified momentarily.

A crucial quantity that appears repeatedly in analyzing the above terms is $\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\|_\infty$, whose importance was already made apparent in the work Azar et al. (2013). A widely used upper bound on this quantity, originally due to (Azar et al. 2013, Lemma 8), is given by

$$\left\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\right\|_\infty \le \frac{2\log 2}{\gamma(1-\gamma)^{1.5}}\|\boldsymbol{r}\|_\infty. \tag{EC.2}$$

This bound turns out to be loose for our purpose, and we develop an improved bound as follows, whose proof is deferred to Appendix EC.2.1.1.

LEMMA EC.2. *Consider any policy π and any probability transition matrix $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$. Let $\boldsymbol{V}$ be a vector obeying $\boldsymbol{V} = (\boldsymbol{I} - \gamma \boldsymbol{P}_\pi)^{-1}\boldsymbol{r}_\pi$ for some $|\mathcal{S}|$-dimensional vector $\boldsymbol{r}_\pi \geq \boldsymbol{0}$. For any $0 < \gamma < 1$, one has*

$$\left\|(\boldsymbol{I} - \gamma \boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\right\|_\infty \leq \frac{4}{\gamma\sqrt{1-\gamma}}\|\boldsymbol{V}\|_\infty. \tag{EC.3}$$

REMARK EC.1. In comparison to the bound (EC.2) derived in (Azar et al. 2013, Lemma 8), Lemma EC.2 offers an improved upper bound stated directly in terms of the properties of $\boldsymbol{V}$ rather than those of $\boldsymbol{r}$.

As it turns out, Lemma EC.2 allows us to obtain an entrywise bound for $\boldsymbol{V}^{(l)}$ $(1 \leq l \leq m)$. To begin with, the first term $\boldsymbol{V}^{(1)}$ satisfies

$$\left\|\boldsymbol{V}^{(1)}\right\|_\infty = \left\|(\boldsymbol{I} - \gamma \boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\left[\boldsymbol{V}^{(0)}\right]}\right\|_\infty \tag{EC.4}$$

since $\boldsymbol{r}^{(1)} = \sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\left[\boldsymbol{V}^{(0)}\right]}$. Next, for any $l > 1$ one has

$$\begin{aligned}\left\|\boldsymbol{V}^{(l)}\right\|_\infty &= \left\|(\boldsymbol{I} - \gamma \boldsymbol{P}_\pi)^{-1}\boldsymbol{r}^{(l)}\right\|_\infty = \left\|(\boldsymbol{I} - \gamma \boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\left[\boldsymbol{V}^{(l-1)}\right]}\right\|_\infty \\ &\leq \frac{4}{\gamma\sqrt{1-\gamma}}\left\|\boldsymbol{V}^{(l-1)}\right\|_\infty,\end{aligned}$$

where the second identity results from the definition of $\boldsymbol{r}^{(l)}$, and the last inequality comes from Lemma EC.2. As a consequence, applying this inequality recursively gives

$$\left\|\boldsymbol{V}^{(l)}\right\|_\infty \leq \left(\frac{4}{\gamma\sqrt{1-\gamma}}\right)^{l-1}\left\|\boldsymbol{V}^{(1)}\right\|_\infty. \tag{EC.5}$$

Main proof. Equipped with the above facts, we are now in a position to prove the lemmas, for which we start with the more general one — Lemma 2. Consider any $0 \leq l \leq m$. We first observe that

$$\begin{aligned}\widehat{\boldsymbol{V}}^{(l)} - \boldsymbol{V}^{(l)} &= (\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_\pi)^{-1}\boldsymbol{r}^{(l)} - \boldsymbol{V}^{(l)} \\ &= \left(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_\pi\right)^{-1}\left(\boldsymbol{I} - \gamma\boldsymbol{P}_\pi\right)\boldsymbol{V}^{(l)} - \left(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_\pi\right)^{-1}\left(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_\pi\right)\boldsymbol{V}^{(l)} \\ &= \gamma\left(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_\pi\right)^{-1}\left(\widehat{\boldsymbol{P}}_\pi - \boldsymbol{P}_\pi\right)\boldsymbol{V}^{(l)},\end{aligned} \tag{EC.6}$$

where the second line follows since, by definition, $\left(\boldsymbol{I} - \gamma\boldsymbol{P}_\pi\right)\boldsymbol{V}^{(l)} = \boldsymbol{r}^{(l)}$. Suppose that there exists some quantity $\beta_1 > 0$ such that the following condition

$$\left|\left(\widehat{\boldsymbol{P}}_\pi - \boldsymbol{P}_\pi\right)\boldsymbol{V}^{(l)}\right| \leq \sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\left[\boldsymbol{V}^{(l)}\right]} + \frac{\left\|\boldsymbol{V}^{(l)}\right\|_\infty\beta_1}{N}\boldsymbol{1} \tag{EC.7}$$

holds uniformly for all $0 \le l \le m$. Then this combined with (EC.6) gives

$$\begin{aligned}\left\|\widehat{\boldsymbol{V}}^{(l)}-\boldsymbol{V}^{(l)}\right\|_\infty &= \gamma\left\|\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\right)^{-1}\left(\widehat{\boldsymbol{P}}_\pi-\boldsymbol{P}_\pi\right)\boldsymbol{V}^{(l)}\right\|_\infty \\ &\overset{\text{(i)}}{\le} \gamma\left\|\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\right)^{-1}\left|\left(\widehat{\boldsymbol{P}}_\pi-\boldsymbol{P}_\pi\right)\boldsymbol{V}^{(l)}\right|\right\|_\infty \\ &\overset{\text{(ii)}}{\le} \gamma\sqrt{\frac{\beta_1}{N}}\left\|\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\right)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\left[\boldsymbol{V}^{(l)}\right]}\right\|_\infty + \frac{\gamma\left\|\boldsymbol{V}^{(l)}\right\|_\infty\beta_1}{N}\left\|\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\right)^{-1}\boldsymbol{1}\right\|_\infty.\end{aligned}$$

Here, (i) follows since $\left(\boldsymbol{I}-\gamma\widehat{\boldsymbol{P}}_\pi\right)^{-1}$ is a non-negative matrix, (ii) comes from (EC.7) and the triangle inequality. Recalling the definition of $\boldsymbol{r}^{(l)}$ and $\widehat{\boldsymbol{V}}^{(l)}$ and invoking Lemma EC.1(d), we can further bound the above as

$$\begin{aligned}\left\|\widehat{\boldsymbol{V}}^{(l)}-\boldsymbol{V}^{(l)}\right\|_\infty &\le \gamma\sqrt{\frac{\beta_1}{N}}\left\|\widehat{\boldsymbol{V}}^{(l+1)}\right\|_\infty + \frac{\gamma\left\|\boldsymbol{V}^{(l)}\right\|_\infty\beta_1}{(1-\gamma)N} \\ &\le \gamma\sqrt{\frac{\beta_1}{N}}\left\|\widehat{\boldsymbol{V}}^{(l+1)}-\boldsymbol{V}^{(l+1)}\right\|_\infty + \gamma\sqrt{\frac{\beta_1}{N}}\left\|\boldsymbol{V}^{(l+1)}\right\|_\infty + \frac{\gamma\beta_1}{(1-\gamma)N}\left\|\boldsymbol{V}^{(l)}\right\|_\infty, \qquad \text{(EC.8)}\end{aligned}$$

where the last inequality is a consequence of the triangle inequality.

The above inequality (EC.8) provides a recursive relation that in turn allows for an effective upper bound. Specifically, combining the inequalities (EC.5) and (EC.8) leads to

$$\begin{aligned}\left\|\widehat{\boldsymbol{V}}^{(0)}-\boldsymbol{V}^{(0)}\right\|_\infty &\le \gamma\sqrt{\frac{\beta_1}{N}}\left\|\widehat{\boldsymbol{V}}^{(1)}-\boldsymbol{V}^{(1)}\right\|_\infty + \gamma\sqrt{\frac{\beta_1}{N}}\left\|\boldsymbol{V}^{(1)}\right\|_\infty + \frac{\gamma\beta_1}{(1-\gamma)N}\left\|\boldsymbol{V}^{(0)}\right\|_\infty \\ &=: b_1\left\|\widehat{\boldsymbol{V}}^{(1)}-\boldsymbol{V}^{(1)}\right\|_\infty + b_1\left\|\boldsymbol{V}^{(1)}\right\|_\infty + \frac{\gamma\beta_1}{(1-\gamma)N}\left\|\boldsymbol{V}^{(0)}\right\|_\infty,\end{aligned}$$

and for $l \ge 1$,

$$\begin{aligned}\left\|\widehat{\boldsymbol{V}}^{(l)}-\boldsymbol{V}^{(l)}\right\|_\infty &\le \gamma\sqrt{\frac{\beta_1}{N}}\left\|\widehat{\boldsymbol{V}}^{(l+1)}-\boldsymbol{V}^{(l+1)}\right\|_\infty + \left(4\sqrt{\frac{\beta_1}{(1-\gamma)N}} + \frac{\gamma\beta_1}{(1-\gamma)N}\right)\left(\frac{4}{\gamma\sqrt{1-\gamma}}\right)^{l-1}\left\|\boldsymbol{V}^{(1)}\right\|_\infty \\ &=: b_1\left\|\widehat{\boldsymbol{V}}^{(l+1)}-\boldsymbol{V}^{(l+1)}\right\|_\infty + b_2 b_3^{l-1}\left\|\boldsymbol{V}^{(1)}\right\|_\infty.\end{aligned}$$

Here for notational simplicity, we introduce

$$b_1 := \gamma\sqrt{\frac{\beta_1}{N}}, \qquad b_2 := 4\sqrt{\frac{\beta_1}{(1-\gamma)N}} + \frac{\gamma\beta_1}{(1-\gamma)N}, \qquad b_3 := \frac{4}{\gamma\sqrt{1-\gamma}}.$$

Invoking the above recursive relation, we can arrange terms to reach

$$\left\|\widehat{\boldsymbol{V}}^{(0)}-\boldsymbol{V}^{(0)}\right\|_\infty \le \underbrace{b_1^m\left\|\widehat{\boldsymbol{V}}^{(m)}-\boldsymbol{V}^{(m)}\right\|_\infty}_{=:\alpha_1} + \underbrace{\left(b_1 b_2\sum_{l=0}^{m-2}(b_1 b_3)^l + b_1\right)}_{=:\alpha_2}\left\|\boldsymbol{V}^{(1)}\right\|_\infty + \frac{\gamma\beta_1}{(1-\gamma)N}\left\|\boldsymbol{V}^{(0)}\right\|_\infty. \qquad \text{(EC.9)}$$

Controlling the quantity α_2. Now it suffices to control the two terms on the right-hand side of the inequality (EC.9) separately, towards which we shall start with the quantity α_2. Assuming that

$N \geq 64\beta_1/(1-\gamma)$, one can easily verify $b_1 b_3 \leq 1/2$. The summation of the geometric sequence thus gives

$$\begin{aligned}\alpha_2 &:= b_1 b_2 \sum_{l=0}^{m-2} (b_1 b_3)^l + b_1 \leq \frac{b_1 b_2}{1-b_1 b_3} + b_1 \leq 2b_1 b_2 + b_1 \\ &= \gamma\sqrt{\frac{\beta_1}{N}}\left\{1 + 8\sqrt{\frac{\beta_1}{(1-\gamma)N}} + \frac{2\gamma\beta_1}{(1-\gamma)N}\right\} \leq 3\gamma\sqrt{\frac{\beta_1}{N}}, \end{aligned} \tag{EC.10}$$

where the last step holds with the assumption $N \geq \frac{64\beta_1}{1-\gamma}$.

Controlling the quantity α_1. Next, we proceed to the quantity α_1, which requires the control of $\big\|\widehat{\boldsymbol{V}}^{(m)} - \boldsymbol{V}^{(m)}\big\|_\infty$. In view of the identity (EC.6), we obtain

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}^{(m)} - \boldsymbol{V}^{(m)}\big\|_\infty &= \gamma\big\|(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_\pi)^{-1}(\boldsymbol{P}_\pi - \widehat{\boldsymbol{P}}_\pi)\boldsymbol{V}^{(m)}\big\|_\infty \\ &\leq \gamma\left\|(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_\pi)^{-1}\left(\sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^{(m)}\big]} + \frac{\big\|\boldsymbol{V}^{(m)}\big\|_\infty \beta_1}{N}\mathbf{1}\right)\right\|_\infty,\end{aligned}$$

where the last inequality follows from the Bernstein-type condition (EC.7) and the fact that $\big(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_\pi\big)^{-1}$ has non-negative entries. By virtue of the simple relation $\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^{(m)}\big]} \leq \big\|\boldsymbol{V}^{(m)}\big\|_\infty$ and the fact that $\big\|\big(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_\pi\big)^{-1}\big\|_1 \leq \frac{1}{1-\gamma}$ (cf. Lemma EC.1(c)), it is further guaranteed that

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}^{(m)} - \boldsymbol{V}^{(m)}\big\|_\infty &\leq \gamma\|(\boldsymbol{I} - \gamma\widehat{\boldsymbol{P}}_\pi)^{-1}\|_1 \cdot \left\|\sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}\big[\boldsymbol{V}^{(m)}\big]} + \frac{\big\|\boldsymbol{V}^{(m)}\big\|_\infty \beta_1}{N}\mathbf{1}\right\|_\infty, \\ &\leq \frac{\gamma}{1-\gamma}\left(\sqrt{\frac{\beta_1}{N}} + \frac{\beta_1}{N}\right)\big\|\boldsymbol{V}^{(m)}\big\|_\infty,\end{aligned}$$

which combined with the bound (EC.5) yields

$$\big\|\widehat{\boldsymbol{V}}^{(m)} - \boldsymbol{V}^{(m)}\big\|_\infty \leq \frac{\gamma}{1-\gamma}\left(\sqrt{\frac{\beta_1}{N}} + \frac{\beta_1}{N}\right)\Big(\frac{4}{\gamma\sqrt{1-\gamma}}\Big)^{m-1}\|\boldsymbol{V}^{(1)}\|_\infty.$$

Putting the above bounds together yields

$$\begin{aligned}\alpha_1 := b_1^m\big\|\widehat{\boldsymbol{V}}^{(m)} - \boldsymbol{V}^{(m)}\big\|_\infty &\leq \left(\gamma\sqrt{\frac{\beta_1}{N}}\right)^m \frac{\gamma}{1-\gamma}\left(\sqrt{\frac{\beta_1}{N}} + \frac{\beta_1}{N}\right)\Big(\frac{4}{\gamma\sqrt{1-\gamma}}\Big)^{m-1}\|\boldsymbol{V}^{(1)}\|_\infty \\ &= \left(\sqrt{\frac{16\beta_1}{N(1-\gamma)}}\right)^{m-1}\left(\sqrt{\frac{\beta_1}{N}} + 1\right)\frac{\gamma^2\beta_1}{(1-\gamma)N}\|\boldsymbol{V}^{(1)}\|_\infty \\ &\leq \left(\frac{1}{e}\right)^{m-1}\frac{1.1\gamma^2\beta_1}{(1-\gamma)N}\|\boldsymbol{V}^{(1)}\|_\infty,\end{aligned} \tag{EC.11}$$

where the last inequality holds provided that $N > \frac{16e^2}{1-\gamma}\beta_1$.

Putting all this together. Combining the inequalities (EC.9), (EC.10) and (EC.11) gives

$$\big\|\widehat{\boldsymbol{V}}^{(0)} - \boldsymbol{V}^{(0)}\big\|_\infty \leq \frac{\gamma\beta_1}{(1-\gamma)N}\big\|\boldsymbol{V}^{(0)}\big\|_\infty + \left\{3\gamma\sqrt{\frac{\beta_1}{N}} + \left(\frac{1}{e}\right)^{m-1}\frac{1.1\gamma^2\beta_1}{(1-\gamma)N}\right\}\|\boldsymbol{V}^{(1)}\|_\infty.$$

To finish up, set $m = \log(\frac{e}{1-\gamma})$ and assume that $N > \frac{16e^2}{1-\gamma}\beta_1$. Recognizing that $\boldsymbol{V}^{(0)} = \boldsymbol{V}^{\pi}$ and $\widehat{\boldsymbol{V}}^{(0)} = \widehat{\boldsymbol{V}}^{\pi}$, we arrive at

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}^{\pi} - \boldsymbol{V}^{\pi}\big\|_{\infty} = \big\|\widehat{\boldsymbol{V}}^{(0)} - \boldsymbol{V}^{(0)}\big\|_{\infty} &\le \frac{\gamma\beta_1}{(1-\gamma)N}\big\|\boldsymbol{V}^{(0)}\big\|_{\infty} + \left\{3\gamma\sqrt{\frac{\beta_1}{N}} + \left(\frac{1}{e}\right)^{m-1}\frac{1.1\gamma^2\beta_1}{(1-\gamma)N}\right\}\|\boldsymbol{V}^{(1)}\|_{\infty} \\ &\le \frac{\gamma\beta_1}{(1-\gamma)N}\big\|\boldsymbol{V}^{\pi}\big\|_{\infty} + 4\gamma\sqrt{\frac{\beta_1}{N}}\Big\|(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\pi}}\left[\boldsymbol{V}^{\pi}\right]}\Big\|_{\infty},\end{aligned} \tag{EC.12}$$

provided that $N \ge \frac{16e^2\beta_1}{1-\gamma}$.

Proof of Lemma 2. Invoking the inequality (EC.2) to bound the second term of (EC.12), we reach

$$\big\|\widehat{\boldsymbol{V}}^{\pi} - \boldsymbol{V}^{\pi}\big\|_{\infty} \le \frac{\gamma\beta_1}{(1-\gamma)N}\big\|\boldsymbol{V}^{\pi}\big\|_{\infty} + 8\log 2\sqrt{\frac{\beta_1}{N(1-\gamma)^3}}\|\boldsymbol{r}\|_{\infty} \le 6\sqrt{\frac{\beta_1}{N(1-\gamma)^3}}\|\boldsymbol{r}\|_{\infty}, \tag{EC.13}$$

where the last inequality uses the elementary fact that $\big\|\boldsymbol{V}^{\pi}\big\|_{\infty} \le \frac{1}{1-\gamma}\|\boldsymbol{r}\|_{\infty}$ and the assumption that $N > \frac{16e^2}{1-\gamma}\beta_1$. We complete the proof of Lemma 2.

Proof of Lemma 1. Finally, to establish Lemma 1, we observe that: for any fixed policy π, the vector $\boldsymbol{V}^{(l)}$ is independent of $\widehat{\boldsymbol{P}}_{\pi}$. The Bernstein inequality (e.g. (Agarwal et al. 2020, Lemma 6)) then reveals that with probability at least $1-\delta$,

$$\Big|\big(\widehat{\boldsymbol{P}}_{\pi} - \boldsymbol{P}_{\pi}\big)\boldsymbol{V}^{(l)}\Big| \le \sqrt{\frac{2\log\left(\frac{4m|\mathcal{S}|}{\delta}\right)}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\pi}}\left[\boldsymbol{V}^{(l)}\right]} + \frac{\big\|\boldsymbol{V}^{(l)}\big\|_{\infty}\log\left(\frac{4m|\mathcal{S}|}{\delta}\right)}{N}\mathbf{1} \tag{EC.14}$$

holds uniformly for all $0 \le l \le m$. This means that we can take $\beta_1 := 2\log\left(\frac{4m|\mathcal{S}|}{\delta}\right)$ with $m = \log(\frac{e}{1-\gamma})$ for this case. Combining this with the inequality (EC.12), we derive the advertised instance-dependent bound

$$\big\|\widehat{\boldsymbol{V}}^{\pi} - \boldsymbol{V}^{\pi}\big\|_{\infty} \le \frac{2\gamma\log\left(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\right)}{(1-\gamma)N}\big\|\boldsymbol{V}^{\pi}\big\|_{\infty} + 4\gamma\sqrt{\frac{2\log\left(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\right)}{(1-\gamma)N}}\Big\|(\boldsymbol{I}-\gamma\boldsymbol{P}_{\pi})^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\pi}}\left[\boldsymbol{V}^{\pi}\right]}\Big\|_{\infty}.$$

Further, this taken collectively with (EC.2) and the crude bound $\|\boldsymbol{V}^{\pi}\|_{\infty} \le \frac{1}{1-\gamma}\|\boldsymbol{r}\|_{\infty}$ gives

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}^{\pi} - \boldsymbol{V}^{\pi}\big\|_{\infty} &\le \frac{2\gamma\log\left(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\right)}{(1-\gamma)^2N}\big\|\boldsymbol{r}\big\|_{\infty} + 8\log 2\sqrt{\frac{2\log\left(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\right)}{(1-\gamma)^3N}}\|\boldsymbol{r}\|_{\infty} \\ &\le 6\sqrt{\frac{2\log\left(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\right)}{(1-\gamma)^3N}}\|\boldsymbol{r}\|_{\infty},\end{aligned}$$

with the proviso that $N \ge \frac{32e^2}{1-\gamma}\log\left(\frac{4|\mathcal{S}|\log\frac{e}{1-\gamma}}{\delta}\right)$.

EC.2.1.1. Proof of Lemma EC.2 To begin with, we make the observation that

$$\begin{aligned}
\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V}) &= \boldsymbol{P}_\pi(\boldsymbol{V}\circ\boldsymbol{V}) - (\boldsymbol{P}_\pi\boldsymbol{V})\circ(\boldsymbol{P}_\pi\boldsymbol{V}) \\
&= \boldsymbol{P}_\pi(\boldsymbol{V}\circ\boldsymbol{V}) - \frac{1}{\gamma^2}(\boldsymbol{V}-\boldsymbol{r}_\pi)\circ(\boldsymbol{V}-\boldsymbol{r}_\pi) \qquad \text{(EC.15)} \\
&= \boldsymbol{P}_\pi(\boldsymbol{V}\circ\boldsymbol{V}) - \frac{1}{\gamma^2}\boldsymbol{V}\circ\boldsymbol{V} + \frac{2}{\gamma^2}\boldsymbol{V}\circ\boldsymbol{r}_\pi - \frac{1}{\gamma^2}\boldsymbol{r}_\pi\circ\boldsymbol{r}_\pi \\
&\le \frac{1}{\gamma^2}(\gamma^2\boldsymbol{P}_\pi - \boldsymbol{I})(\boldsymbol{V}\circ\boldsymbol{V}) + \frac{2}{\gamma^2}\boldsymbol{V}\circ\boldsymbol{r}_\pi, \qquad \text{(EC.16)}
\end{aligned}$$

where the identity (EC.15) makes use of the relation $\boldsymbol{V} = \boldsymbol{r}_\pi + \gamma\boldsymbol{P}_\pi\boldsymbol{V}$. In addition, one can deduce that

$$\begin{aligned}
\left\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\right\|_\infty &= \frac{1}{1-\gamma}\left\|(1-\gamma)(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\right\|_\infty \\
&\overset{\text{(i)}}{\le} \frac{1}{\sqrt{1-\gamma}}\left\|\sqrt{(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\right\|_\infty \\
&\overset{\text{(ii)}}{\le} \frac{1}{\sqrt{1-\gamma}}\sqrt{\|2(\boldsymbol{I}-\gamma^2\boldsymbol{P}_\pi)^{-1}\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})\|_\infty}.
\end{aligned}$$

Here, (i) comes from Jensen's inequality (so that $\mathbb{E}[\sqrt{v}] \le \sqrt{\mathbb{E}[v]}$) recognizing that each row of $(1-\gamma)(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}$ is a probability distribution, and Lemma EC.1(d), (ii) is an elementary fact established in (Agarwal et al. 2020, Lemma 4). Combining Lemma EC.1(e) and the inequality (EC.16) further yields

$$\begin{aligned}
\left\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\right\|_\infty &\le \frac{1}{\sqrt{1-\gamma}}\sqrt{\left\|2(\boldsymbol{I}-\gamma^2\boldsymbol{P}_\pi)^{-1}\left(\frac{1}{\gamma^2}\left(\gamma^2\boldsymbol{P}_\pi-\boldsymbol{I}\right)(\boldsymbol{V}\circ\boldsymbol{V}) + \frac{2}{\gamma^2}\boldsymbol{V}\circ\boldsymbol{r}_\pi\right)\right\|_\infty} \\
&\le \frac{1}{\gamma\sqrt{1-\gamma}}\sqrt{2\|\boldsymbol{V}\circ\boldsymbol{V}\|_\infty} + \frac{2}{\gamma\sqrt{1-\gamma}}\sqrt{\|(\boldsymbol{I}-\gamma^2\boldsymbol{P}_\pi)^{-1}(\boldsymbol{V}\circ\boldsymbol{r}_\pi)\|_\infty}. \qquad \text{(EC.17)}
\end{aligned}$$

where the last step arises from the triangle inequality and $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$. This leaves us with two terms to deal with.

Regarding the first term of (EC.17), we observe that $\|(\boldsymbol{V}\circ\boldsymbol{V})\|_\infty = \|\boldsymbol{V}\|_\infty^2$. When it comes to the second term of (EC.17), it is seen that

$$\begin{aligned}
\left\|(\boldsymbol{I}-\gamma^2\boldsymbol{P}_\pi)^{-1}(\boldsymbol{V}\circ\boldsymbol{r}_\pi)\right\|_\infty &\le \left\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}(\boldsymbol{V}\circ\boldsymbol{r}_\pi)\right\|_\infty \\
&\le \|\boldsymbol{V}\|_\infty\left\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\boldsymbol{r}_\pi\right\|_\infty \\
&= \|\boldsymbol{V}\|_\infty^2.
\end{aligned}$$

Here, the first inequality holds true since $(\boldsymbol{I}-\gamma^2\boldsymbol{P}_\pi)^{-1} = \sum_{i=0}^\infty \gamma^{2i}\boldsymbol{P}_\pi^i \le \sum_{i=0}^\infty \gamma^i\boldsymbol{P}_\pi^i = (\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}$, while the second line holds since $\boldsymbol{V}$, $\boldsymbol{r}$ and $(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}$ are all non-negative. Substitution into (EC.17) thus yields

$$\left\|(\boldsymbol{I}-\gamma\boldsymbol{P}_\pi)^{-1}\sqrt{\mathsf{Var}_{\boldsymbol{P}_\pi}(\boldsymbol{V})}\right\|_\infty \le \frac{\sqrt{2}}{\gamma\sqrt{1-\gamma}}\|\boldsymbol{V}\|_\infty + \frac{2}{\gamma\sqrt{1-\gamma}}\|\boldsymbol{V}\|_\infty \le \frac{4}{\gamma\sqrt{1-\gamma}}\|\boldsymbol{V}\|_\infty$$

as claimed.

EC.2.2. Proof of Lemma 3

To establish Lemma 3, it suffices to check that $\boldsymbol{V}^\star$ and $\boldsymbol{Q}^\star$ satisfy the Bellman optimality equations underlying $\mathcal{M}_{s,a,u^\star}$. Towards this end, we study the absorbing state-action pair (s,a) and other pairs separately. For notational simplicity, we shall let $\boldsymbol{P}^{\mathsf{abs}}$ and $r^{\mathsf{abs}}(\cdot,\cdot)$ denote respectively the probability transition matrix and the reward function associated with $\mathcal{M}_{s,a,u^\star}$.

First, we observe that, by construction,

$$r^{\mathsf{abs}}(s,a)+\gamma(\boldsymbol{P}^{\mathsf{abs}}\boldsymbol{V}^\star)_{s,a}=u^\star+\gamma V^\star(s).$$

Recall that $\boldsymbol{V}^\star$ satisfies the Bellman optimality equation w.r.t. the original MDP, namely, $Q^\star(s,a)=r(s,a)+\gamma(\boldsymbol{P}\boldsymbol{V}^\star)_{s,a}$. This together with our choice of $u^\star$ gives

$$u^\star+\gamma V^\star(s)=Q^\star(s,a)-\gamma V^\star(s)+\gamma V^\star(s)=Q^\star(s,a).$$

Putting the above identities together, we arrive at

$$Q^\star(s,a)=r^{\mathsf{abs}}(s,a)+\gamma(\boldsymbol{P}^{\mathsf{abs}}\boldsymbol{V}^\star)_{s,a}. \tag{EC.18}$$

Next, consider any state-action pair $(s',a')\neq(s,a)$. Recalling again the properties of $\mathcal{M}_{s,a,u^\star}$, we reach

$$r^{\mathsf{abs}}(s',a')+\gamma(\boldsymbol{P}^{\mathsf{abs}}\boldsymbol{V}^\star)_{s',a'}=r(s',a')+\gamma(\boldsymbol{P}\boldsymbol{V}^\star)_{s',a'}=Q^\star(s',a'). \tag{EC.19}$$

Here the last identity is due to the Bellman equation w.r.t. the original MDP. Combining (EC.18) and (EC.19) implies that $\boldsymbol{V}^\star$ and $\boldsymbol{Q}^\star$ satisfy Bellman's optimality equations in $\mathcal{M}_{s,a,u^\star}$, thus concluding the proof.

EC.2.3. Proof of Lemma 4

Our first observation is that $\widehat{\boldsymbol{Q}}^\star_{s,a,u}$ satisfies Lipschitz continuity w.r.t. u in the sense that

$$\big\|\widehat{\boldsymbol{Q}}^\star_{s,a,u}-\widehat{\boldsymbol{Q}}^\star_{s,a,u'}\big\|_\infty\le\frac{1}{1-\gamma}|u-u'|. \tag{EC.20}$$

The proof of this relation is identical to that of (Agarwal et al. 2020, Lemma 8); we omit here for brevity. In view of Lemma 3, if we set $\widehat{u}^\star:=r(s,a)+\gamma(\widehat{\boldsymbol{P}}\widehat{\boldsymbol{V}}^\star)_{s,a}-\gamma\widehat{V}^\star(s)$, then one has

$$\widehat{\boldsymbol{Q}}^\star_{s,a,\widehat{u}^\star}=\widehat{\boldsymbol{Q}}^\star\qquad\text{and}\quad\widehat{\boldsymbol{V}}^\star_{s,a,\widehat{u}^\star}=\widehat{\boldsymbol{V}}^\star.$$

In addition, there exists a point u_0 in the epsilon-net $\mathcal{N}_{(1-\gamma)\omega/4}$ such that $|\widehat{u}^\star-u_0|\le(1-\gamma)\omega/4$, which combined with the Lipschitz continuity property (EC.20) gives

$$\big\|\widehat{\boldsymbol{Q}}^\star-\widehat{\boldsymbol{Q}}^\star_{s,a,u_0}\big\|_\infty=\big\|\widehat{\boldsymbol{Q}}^\star_{s,a,\widehat{u}^\star}-\widehat{\boldsymbol{Q}}^\star_{s,a,u_0}\big\|_\infty\le\frac{1}{1-\gamma}|\widehat{u}^\star-u_0|\le\frac{\omega}{4}. \tag{EC.21}$$

Additionally, for any $s' \in \mathcal{S}$ and any $a_1, a_2 \in \mathcal{A}$ with $a_1 \neq a_2$, we have the following decomposition

$$\begin{aligned}
&\widehat{Q}^\star_{s,a,u_0}(s',a_1) - \widehat{Q}^\star_{s,a,u_0}(s',a_2) \\
&\quad = \widehat{Q}^\star(s',a_1) - \widehat{Q}^\star(s',a_2) + \widehat{Q}^\star_{s,a,u_0}(s',a_1) - \widehat{Q}^\star(s',a_1) - \Big(\widehat{Q}^\star_{s,a,u_0}(s',a_2) - \widehat{Q}^\star(s',a_2)\Big) \\
&\quad \geq \widehat{Q}^\star(s',a_1) - \widehat{Q}^\star(s',a_2) - 2\big\|\widehat{\boldsymbol{Q}}^\star - \widehat{\boldsymbol{Q}}^\star_{s,a,u_0}\big\|_\infty \\
&\quad \geq \widehat{Q}^\star(s',a_1) - \widehat{Q}^\star(s',a_2) - \frac{\omega}{2},
\end{aligned} \tag{EC.22}$$

where the last inequality invokes the inequality (EC.21). Moreover, our separation condition defined in (41) requires that: for any $s' \in \mathcal{S}$ and any $a_2 \neq \widehat{\pi}^\star(s')$, one has $\widehat{Q}^\star(s', \widehat{\pi}^\star(s')) - \widehat{Q}^\star(s',a_2) \geq \omega$, which together with (EC.22) reveals that

$$\widehat{Q}^\star_{s,a,u_0}(s', \widehat{\pi}^\star(s')) - \widehat{Q}^\star_{s,a,u_0}(s',a_2) \geq \widehat{Q}^\star(s', \widehat{\pi}^\star(s')) - \widehat{Q}^\star(s',a_2) - \frac{\omega}{2} \geq \frac{\omega}{2}. \tag{EC.23}$$

Given that $\widehat{\pi}^\star_{s,a,u_0}(s') := \arg\max_{a'} \widehat{Q}^\star_{s,a,u_0}(s',a')$, it is seen from (EC.23) that

$$\widehat{\pi}^\star_{s,a,u_0}(s') = \widehat{\pi}^\star(s'),$$

which holds true for all $s' \in \mathcal{S}$. This concludes the proof.

EC.2.4. Proof of Lemma 5

We start by bounding $\big\|\widehat{\boldsymbol{V}}^{\widehat{\pi}^\star} - \boldsymbol{V}^{\widehat{\pi}^\star}\big\|_\infty$. Recall the definition of the series $\{\boldsymbol{V}^{(l)}\}$ in (EC.1). Throughout this proof, we shall write $\boldsymbol{V}^{(l)}_\pi$ instead in order to make apparent the dependency on the policy π.

For each state-action pair (s,a), let us construct the epsilon-net $\mathcal{N}_{(1-\gamma)\omega/4}$ as in the expression (42). For every $u \in \mathcal{N}_{(1-\gamma)\omega/4}$, recall that $\widehat{\pi}^\star_{s,a,u}$ is defined as the optimal policy with respect to the (s,a)-absorbing MDP $\widehat{\mathcal{M}}_{s,a,u}$. By construction, the set of policies $\widehat{\pi}^\star_{s,a,u}$ $(u \in \mathcal{N}_{(1-\gamma)\omega/4})$ is independent of $\widehat{\boldsymbol{P}}_{s,a}$. The Bernstein inequality (e.g. (Agarwal et al. 2020, Lemma 6)) taken together with the union bound thus guarantees that with probability at least $1-\delta$,

$$\Big|(\widehat{\boldsymbol{P}} - \boldsymbol{P})_{s,a} \boldsymbol{V}^{(l)}_{\widehat{\pi}^\star_{s,a,u}}\Big| \leq \sqrt{\frac{\beta_1}{N}} \sqrt{\big(\mathsf{Var}_{\boldsymbol{P}}\big[\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star_{s,a,u}}\big]\big)_{s,a}} + \frac{\big\|\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star_{s,a,u}}\big\|_\infty \beta_1}{N} \tag{EC.24}$$

holds uniformly over all $0 \leq l \leq \log \frac{e}{1-\gamma}$, $u \in \mathcal{N}_{(1-\gamma)\omega/4}$, $(s,a) \in \mathcal{S} \times \mathcal{A}$. Here, β_1 is given by

$$\beta_1 := 2\log\Big(\frac{4\log\big(\frac{e}{1-\gamma}\big)|\mathcal{N}_{(1-\gamma)\omega/4}||\mathcal{S}||\mathcal{A}|}{\delta}\Big) \leq 2\log\Big(\frac{32}{(1-\gamma)^2\omega\delta}|\mathcal{S}||\mathcal{A}|\log\big(\frac{e}{1-\gamma}\big)\Big),$$

where we have used the fact $|\mathcal{N}_{(1-\gamma)\omega/4}| \leq \frac{8}{(1-\gamma)^2\omega}$. In addition, for any $0 < \omega < 1$, Lemma 4 guarantees that for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, there exists a point $u_0 \in \mathcal{N}_{(1-\gamma)\omega/4}$ such that $\widehat{\pi}^\star = \widehat{\pi}^\star_{s,a,u_0}$. Invoking this important fact, we obtain

$$\Big|(\widehat{\boldsymbol{P}} - \boldsymbol{P})_{s,a} \boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\Big| = \Big|(\widehat{\boldsymbol{P}} - \boldsymbol{P})_{s,a} \boldsymbol{V}^{(l)}_{\widehat{\pi}^\star_{s,a,u_0}}\Big|$$

$$\leq \sqrt{\frac{\beta_1}{N}}\sqrt{\left(\mathsf{Var}_{\boldsymbol{P}}\left[\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star_{s,a,u_0}}\right]\right)_{s,a}} + \frac{\left\|\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star_{s,a,u_0}}\right\|_\infty \beta_1}{N}$$
$$= \sqrt{\frac{\beta_1}{N}}\sqrt{\left(\mathsf{Var}_{\boldsymbol{P}}\left[\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\right]\right)_{s,a}} + \frac{\left\|\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\right\|_\infty \beta_1}{N}.$$

The above inequality further allows us to deduce that, with probability $1-\delta$,

$$\left|\left(\widehat{\boldsymbol{P}}_{\widehat{\pi}^\star} - \boldsymbol{P}_{\widehat{\pi}^\star}\right)\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\right| = \left|\boldsymbol{\Pi}^{\widehat{\pi}^\star}(\widehat{\boldsymbol{P}}-\boldsymbol{P})\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\right|$$
$$\leq \sqrt{\frac{\beta_1}{N}}\sqrt{\boldsymbol{\Pi}^{\widehat{\pi}^\star}\mathsf{Var}_{\boldsymbol{P}}\left[\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\right]} + \frac{\left\|\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\right\|_\infty \beta_1}{N}\mathbf{1}$$
$$= \sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{\widehat{\pi}^\star}}\left[\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\right]} + \frac{\left\|\boldsymbol{V}^{(l)}_{\widehat{\pi}^\star}\right\|_\infty \beta_1}{N}\mathbf{1}.$$

The above derivation validates the assumption required for Lemma 2. As a result, if $N > \frac{16e^2}{1-\gamma}\beta_1$ and $0<\omega<1$, then Lemma 2 leads to the advertised bound

$$\left\|\widehat{\boldsymbol{V}}^{\widehat{\pi}^\star} - \boldsymbol{V}^{\widehat{\pi}^\star}\right\|_\infty \leq \frac{6}{1-\gamma}\sqrt{\frac{\beta_1}{N(1-\gamma)}} \leq \frac{6}{1-\gamma}\sqrt{\frac{2\log\left(\frac{32}{(1-\gamma)^2\omega\delta}|\mathcal{S}||\mathcal{A}|\log\left(\frac{e}{1-\gamma}\right)\right)}{N(1-\gamma)}}$$
$$\leq 6\sqrt{\frac{2\log\left(\frac{32|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\omega\delta}\right)}{N(1-\gamma)^3}}. \tag{EC.25}$$

Finally, we move on to the term $\boldsymbol{V}^\star - \widehat{\boldsymbol{V}}^{\pi^\star}$. Given that $\pi^\star$ is independent of $\widehat{\boldsymbol{P}}$, invoke Lemma 1 to reach

$$\left\|\widehat{\boldsymbol{V}}^{\pi^\star} - \boldsymbol{V}^\star\right\|_\infty = \left\|\widehat{\boldsymbol{V}}^{\pi^\star} - \boldsymbol{V}^{\pi^\star}\right\|_\infty \leq 6\sqrt{2}\sqrt{\frac{\log\left(\frac{4|\mathcal{S}|}{\delta}\right)+\log\log\left(\frac{e}{1-\gamma}\right)}{N(1-\gamma)^3}}$$
$$\leq 6\sqrt{\frac{2\log\left(\frac{32|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2\omega\delta}\right)}{N(1-\gamma)^3}} \tag{EC.26}$$

with probability at least $1-\delta$. This together with (34) and (EC.25) immediately establishes Lemma 5.

EC.2.5. Proof of Lemma 6

The proofs for $Q^\star_{\mathrm{p}}$ and $\widehat{Q}^\star_{\mathrm{p}}$ are exactly the same; for the sake of conciseness, we shall only provide the proof for $Q^\star_{\mathrm{p}}$. Here we aim to prove a more general result than Lemma 6, namely, with probability at least $1-\delta$,

$$\forall s\in\mathcal{S} \text{ and } a_1,a_2\in\mathcal{A} \text{ with } a_1\neq a_2: \qquad \left|Q^\star_{\mathrm{p}}(s,a_1) - Q^\star_{\mathrm{p}}(s,a_2)\right| > \frac{\xi\delta(1-\gamma)}{4|\mathcal{S}||\mathcal{A}|^2}.$$

Consider any state s and any actions $a_1\neq a_2$. In what follows, we allow $r_{\mathrm{p}}(s,a_1)=\tau$ to vary, while *freezing* the values of all other rewards $\{r_{\mathrm{p}}(\tilde{s},a)\,|\,(\tilde{s},a)\neq(s,a_1)\}$. To streamline notation, we define

- $\boldsymbol{r}_\tau = [r_\tau(s,a)]_{(s,a)\in\mathcal{S}\times\mathcal{A}}$: the reward vector obeying

$$r_\tau(s,a_1) = \tau \qquad \text{and} \qquad r_\tau(\tilde{s},a) = r_{\mathrm{p}}(\tilde{s},a) \quad \text{for all } (\tilde{s},a)\neq(s,a_1);$$

- $Q_\tau^\star$: the optimal Q-function when the reward vector is $\boldsymbol{r}_\tau$;
- $V_\tau^\star$: the optimal value function when the reward vector is $\boldsymbol{r}_\tau$;
- $\pi_\tau^\star$: the optimal policy when the reward vector is $\boldsymbol{r}_\tau$.

Additionally, we claim for the moment that there exists a phase transition boundary τ_{th} such that

$$\pi_\tau^\star(s) \neq a_1, \quad \text{for all } \tau < \tau_{\mathrm{th}}; \tag{EC.27a}$$

$$\pi_\tau^\star(s) = a_1, \quad \text{for all } \tau > \tau_{\mathrm{th}}. \tag{EC.27b}$$

The proof of this claim is deferred to the end of this section. To establish Lemma 6, the idea is to control the size of the set

$$\mathcal{I}_{0,\omega} := \left\{\tau \mid \left|Q_\tau^\star(s,a_1) - Q_\tau^\star(s,a_2)\right| < \omega\right\} \tag{EC.28}$$

for some $\omega > 0$ to be specified shortly. As motivated by (EC.27), we further break down this set into two parts $\mathcal{I}_{0,\omega} = \mathcal{I}_{1,\omega} \cup \mathcal{I}_{2,\omega}$, where

$$\mathcal{I}_{1,\omega} := \left\{\tau \mid \tau < \tau_{\mathrm{th}},\ \left|Q_\tau^\star(s,a_1) - Q_\tau^\star(s,a_2)\right| < \omega\right\}, \tag{EC.29a}$$

$$\mathcal{I}_{2,\omega} := \left\{\tau \mid \tau \geq \tau_{\mathrm{th}},\ \left|Q_\tau^\star(s,a_1) - Q_\tau^\star(s,a_2)\right| < \omega\right\}. \tag{EC.29b}$$

In what follows, we first control the size of each set separately, and then demonstrate that the probability of these events happening is very small.

Step 1. We begin with $\mathcal{I}_{1,\omega}$ associated with the range $\tau < \tau_{\mathrm{th}}$. In this case, the value function $V_\tau^\star$ does not vary with τ, since the reward $r_\tau(s,a_1) = \tau$ is never active when calculating $V_\tau^\star$ (by virtue of (EC.27a)). Thus, the Bellman equation allows us to write

$$Q_\tau^\star(s,a_1) = \tau + B_1 \qquad \text{and} \qquad Q_\tau^\star(s,a_2) = B_2$$

for some quantities B_1 and B_2, where neither B_1 nor B_2 depends on the value of τ. Armed with this observation, we can easily show that: for any $\omega > 0$, the interval $\mathcal{I}_{1,\omega}$ (cf. (EC.29a)) obeys

$$\mathcal{I}_{1,\omega} \subseteq \left\{\tau \mid |\tau + B_1 - B_2| < \omega\right\},$$

and hence has length (or Lebesgue measure) at most 2ω.

Step 2. We then move on to $\mathcal{I}_{2,\omega}$ associated with the range $\tau > \tau_{\mathrm{th}}$ in which case $\pi_\tau^\star(s) = a_1$. Towards this, we first make some useful observations.

- To begin with, given the relation $\boldsymbol{Q}_\tau^\star = \boldsymbol{r}_\tau + \gamma \boldsymbol{P}\boldsymbol{V}_\tau^\star$, it is easily seen that for any $\tau_2 > \tau_1 > \tau_{\mathrm{th}}$,

$$\boldsymbol{0} \le \boldsymbol{Q}_{\tau_2}^\star - \boldsymbol{Q}_{\tau_1}^\star \le \boldsymbol{r}_{\tau_2} - \boldsymbol{r}_{\tau_1} + \gamma \|\boldsymbol{V}_{\tau_2}^\star - \boldsymbol{V}_{\tau_1}^\star\|_\infty. \tag{EC.30}$$

In addition, for any state-action pair $(\tilde{s}, a) \neq (s, a_1)$, by construction we have $r_{\tau_2}(\tilde{s}, a) - r_{\tau_1}(\tilde{s}, a) = 0$, which together with (EC.30) indicates that

$$\forall (\tilde{s}, a) \neq (s, a_1): \quad 0 \le Q_{\tau_2}^\star(\tilde{s}, a) - Q_{\tau_1}^\star(\tilde{s}, a) \le \gamma \|\boldsymbol{V}_{\tau_2}^\star - \boldsymbol{V}_{\tau_1}^\star\|_\infty. \tag{EC.31}$$

- Next, observe that for any $\tau_2 > \tau_1 > \tau_{\mathrm{th}}$,

$$\forall s' \in \mathcal{S}: \quad 0 \le V_{\tau_2}^\star(s') - V_{\tau_1}^\star(s') = \max_a Q_{\tau_2}^\star(s', a) - \max_a Q_{\tau_1}^\star(s', a) \le \left\|\boldsymbol{Q}_{\tau_2}^\star - \boldsymbol{Q}_{\tau_1}^\star\right\|_\infty \tag{EC.32}$$

and hence $\|\boldsymbol{Q}_{\tau_2}^\star - \boldsymbol{Q}_{\tau_1}^\star\|_\infty \ge \|\boldsymbol{V}_{\tau_2}^\star - \boldsymbol{V}_{\tau_1}^\star\|_\infty$. This combined with (EC.31) and the fact $\gamma < 1$ implies that

$$Q_{\tau_2}^\star(s, a_1) - Q_{\tau_1}^\star(s, a_1) \ge \|\boldsymbol{V}_{\tau_2}^\star - \boldsymbol{V}_{\tau_1}^\star\|_\infty, \tag{EC.33}$$

which together with the facts $V_{\tau_1}^\star(s) = Q_{\tau_1}^\star(s, a_1)$ and $V_{\tau_2}^\star(s) = Q_{\tau_2}^\star(s, a_1)$ (by virtue of (EC.27b)) yields

$$\begin{aligned} V_{\tau_2}^\star(s) - V_{\tau_1}^\star(s) = Q_{\tau_2}^\star(s, a_1) - Q_{\tau_1}^\star(s, a_1) &\ge \|\boldsymbol{V}_{\tau_2}^\star - \boldsymbol{V}_{\tau_1}^\star\|_\infty \\ \Longrightarrow \quad Q_{\tau_2}^\star(s, a_1) - Q_{\tau_1}^\star(s, a_1) &= \|\boldsymbol{V}_{\tau_2}^\star - \boldsymbol{V}_{\tau_1}^\star\|_\infty. \end{aligned}$$

Invoke the Bellman equation to further derive

$$\begin{aligned} Q_{\tau_2}^\star(s, a_1) - Q_{\tau_1}^\star(s, a_1) = \|\boldsymbol{V}_{\tau_2}^\star - \boldsymbol{V}_{\tau_1}^\star\|_\infty &= \left\|\boldsymbol{r}_{\tau_2} + \gamma \boldsymbol{P}\boldsymbol{V}_{\tau_2}^\star - \boldsymbol{r}_{\tau_1} - \gamma \boldsymbol{P}\boldsymbol{V}_{\tau_1}^\star\right\|_\infty \\ &\ge \left\|\boldsymbol{r}_{\tau_2} - \boldsymbol{r}_{\tau_1}\right\|_\infty = \tau_2 - \tau_1, \end{aligned} \tag{EC.34}$$

where the last inequality holds since $\boldsymbol{r}_{\tau_2} + \gamma \boldsymbol{P}\boldsymbol{V}_{\tau_2}^\star - \boldsymbol{r}_{\tau_1} - \gamma \boldsymbol{P}\boldsymbol{V}_{\tau_1}^\star \ge \boldsymbol{r}_{\tau_2} - \boldsymbol{r}_{\tau_1} \ge \boldsymbol{0}$ (due to the monotonicity properties $\boldsymbol{r}_{\tau_2} \ge \boldsymbol{r}_{\tau_1}$ and $\boldsymbol{V}_{\tau_2}^\star \ge \boldsymbol{V}_{\tau_1}^\star$), and the last identity follows from the definition of $\boldsymbol{r}_\tau$.

With the above two properties (EC.31) and (EC.34) in mind, we are ready to locate $\mathcal{I}_{2,\omega}$ by showing that

$$\mathcal{I}_{2,\omega} \subseteq \left[\tau_{\mathrm{th}}, \tau_{\mathrm{th}} + \frac{\omega}{1-\gamma}\right]. \tag{EC.35}$$

Given that $Q_{\tau_{\mathrm{th}}}^\star(s, a_1) \ge Q_{\tau_{\mathrm{th}}}^\star(s, a_2)$ (in view of (EC.27)), we have for any $\tau \ge \tau_{\mathrm{th}}$ and any $a_2 \neq a_1$ that

$$Q_\tau^\star(s, a_1) - Q_\tau^\star(s, a_2) \ge \left(Q_\tau^\star(s, a_1) - Q_{\tau_{\mathrm{th}}}^\star(s, a_1)\right) - \left(Q_\tau^\star(s, a_2) - Q_{\tau_{\mathrm{th}}}^\star(s, a_2)\right)$$

$$\geq (1-\gamma)\big(Q_{\tau}^{\star}(s,a_1)-Q_{\tau_{\mathrm{th}}}^{\star}(s,a_1)\big). \tag{EC.36}$$

Here, the last inequality holds since

$$Q_{\tau}^{\star}(s,a_2)-Q_{\tau_{\mathrm{th}}}^{\star}(s,a_2)\overset{\text{(i)}}{\leq}\gamma\big\|\boldsymbol{V}_{\tau}^{\star}-\boldsymbol{V}_{\tau_{\mathrm{th}}}^{\star}\big\|_{\infty}\overset{\text{(ii)}}{\leq}\gamma\big(Q_{\tau}^{\star}(s,a_1)-Q_{\tau_{\mathrm{th}}}^{\star}(s,a_1)\big),$$

where (i) follows from (EC.31) and (ii) is due to (EC.33). As a result, for any $\tau>\tau_{\mathrm{th}}+\frac{\omega}{1-\gamma}$, one can invoke (EC.36) and (EC.34) to see that

$$\forall a_2\neq a_1:\quad Q_{\tau}^{\star}(s,a_1)-Q_{\tau}^{\star}(s,a_2)\geq(1-\gamma)\big(Q_{\tau}^{\star}(s,a_1)-Q_{\tau_{\mathrm{th}}}^{\star}(s,a_1)\big)\geq(1-\gamma)(\tau-\tau_{\mathrm{th}})>\omega,$$

which necessarily implies that such a τ does not lie within the interval $\mathcal{I}_{2,\omega}$ as defined in (EC.29b). This establishes the claimed relation (EC.35).

Step 3. Putting the results in the above two steps together, we see the set $\mathcal{I}_{0,\omega}$ (cf. (EC.28)) has total length (or Lebesgue measure) at most $\frac{3\omega}{1-\gamma}$. Given that $r_{\mathrm{p}}(s,a)=r(s,a)+\zeta(s,a)$ with $\zeta(s,a)\sim\mathsf{Unif}(0,\xi)$, one has

$$\mathbb{P}\left\{\left|Q_{\mathrm{p}}^{\star}(s,a_1)-Q_{\mathrm{p}}^{\star}(s,a_2)\right|<\omega\right\}\leq\mathbb{P}\left\{r(s,a)+\zeta(s,a)\in\mathcal{I}_{0,\omega}\right\}\leq\frac{3\omega}{\xi(1-\gamma)}.$$

By setting $\omega=\frac{\delta(1-\gamma)\xi}{3|\mathcal{S}||\mathcal{A}|^2}$, we arrive at

$$\mathbb{P}\left\{\left|Q_{\mathrm{p}}^{\star}(s,a_1)-Q_{\mathrm{p}}^{\star}(s,a_2)\right|<\frac{\delta(1-\gamma)\xi}{3|\mathcal{S}||\mathcal{A}|^2}\right\}\leq\frac{\delta}{|\mathcal{S}||\mathcal{A}|^2}.$$

Finally, taking the union bound over all s,a_1,a_2, we conclude that

$$\mathbb{P}\left\{\exists s,a_1\neq a_2:\ \left|Q_{\mathrm{p}}^{\star}(s,a_1)-Q_{\mathrm{p}}^{\star}(s,a_2)\right|<\frac{\delta(1-\gamma)\xi}{3|\mathcal{S}||\mathcal{A}|^2}\right\}\leq\delta,$$

thus establishing Lemma 6 as long as the claim (EC.27) is valid.

Proof of the claim (EC.27). To establish the claim, it suffices to take

$$\tau_{\mathrm{th}}=\sup\left\{u\,|\,\pi_{\tau}^{\star}(s)\neq a_1\text{ for all }\tau<u\right\}. \tag{EC.37}$$

It thus suffices to verify (EC.27b) for our choice (EC.37). Towards this, suppose instead that there exist some $\tau_3<\tau_{\mathrm{th}}\leq\tau_2<\tau_1$ such that

$$\pi_{\tau_3}^{\star}(s)\neq a_1,\qquad\pi_{\tau_2}^{\star}(s)=a_1,\qquad\text{and}\qquad\pi_{\tau_1}^{\star}(s)\neq a_1.$$

It is straightforward to see that $V_{\tau_1}^{\star}=V_{\tau_3}^{\star}$, since in both cases, the reward $r_{\tau}(s,a_1)$ does not enter the calculation of the optimal value function (while the rewards in other state-action pairs are identical in both cases). In view of the monotonicity of the value function w.r.t. the reward vector, we have

$$V_{\tau_1}^{\star}=V_{\tau_2}^{\star}=V_{\tau_3}^{\star}.$$

However, this contradicts our assumption that a_1 is the optimal action for state s at τ_2 but not at τ_3, since enlarging τ_2 to τ_3 otherwise will enlarge the optimal value function $V_{\tau_3}^{\star}$. We have thus established (EC.27). □

EC.2.6. Proof of Lemma 7

To begin, we find it helpful to introduce a modified reward function $\widetilde{\boldsymbol{r}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as follows

$$\widetilde{r}(s,a) := \begin{cases} r(s,a) + \widehat{V}^\star(s) - \widehat{Q}^\star(s,a), & \text{if } a = \widehat{\pi}_{\mathrm{c}}(s), \\ r(s,a), & \text{otherwise.} \end{cases} \tag{EC.38}$$

Armed with this new reward function, we subsequently define a vector $\widetilde{\boldsymbol{Q}} = [\widetilde{Q}(s,a)]$ as follows

$$\widetilde{\boldsymbol{Q}} := \widetilde{\boldsymbol{r}} + \gamma \widehat{\boldsymbol{P}} \widehat{\boldsymbol{V}}^\star. \tag{EC.39}$$

In view of the Bellman optimality equation $\widehat{\boldsymbol{Q}}^\star = \boldsymbol{r} + \gamma \widehat{\boldsymbol{P}} \widehat{\boldsymbol{V}}^\star$, we see that $\widetilde{\boldsymbol{Q}}$ satisfies $\widetilde{\boldsymbol{Q}} = \widetilde{\boldsymbol{r}} + \widehat{\boldsymbol{Q}}^\star - \boldsymbol{r}$, which combined with the construction (EC.38) gives

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad \widetilde{Q}(s,a) = \begin{cases} \widehat{V}^\star(s), & \text{if } a = \widehat{\pi}_{\mathrm{c}}(s), \\ \widehat{Q}^\star(s,a), & \text{else.} \end{cases}$$

As a consequence, it is easily seen that

$$\forall s \in \mathcal{S}: \qquad \max_a \widetilde{Q}(s,a) = \widehat{V}^\star(s) \qquad \text{and} \qquad \widetilde{Q}\big(s, \widehat{\pi}_{\mathrm{c}}(s)\big) = \widehat{V}^\star(s)$$

This taken collectively with (EC.39) demonstrates that $\widetilde{\boldsymbol{Q}}$ and $\widehat{\boldsymbol{V}}^\star$ are respectively the optimal Q-function and optimal value function of the MDP $\widetilde{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \widetilde{r}, \widehat{P}, \gamma)$, since they satisfy the Bellman optimality condition w.r.t. $\widetilde{\mathcal{M}}$. In addition, if we let $\widetilde{\boldsymbol{V}}^{\widehat{\pi}_{\mathrm{c}}}$ represent the value function of the policy $\widehat{\pi}_{\mathrm{c}}$ in $\widetilde{\mathcal{M}}$, then the preceding relation clearly implies that $\widetilde{\boldsymbol{V}}^{\widehat{\pi}_{\mathrm{c}}} = \widehat{\boldsymbol{V}}^\star$.

Using the above properties, one can deduce that

$$\begin{aligned} \widehat{\boldsymbol{V}}^\star - \widehat{\boldsymbol{V}}^{\widehat{\pi}_{\mathrm{c}}} = \widetilde{\boldsymbol{V}}^{\widehat{\pi}_{\mathrm{c}}} - \widehat{\boldsymbol{V}}^{\widehat{\pi}_{\mathrm{c}}} &\overset{\text{(i)}}{=} \big(\boldsymbol{I} - \gamma \widehat{\boldsymbol{P}}_{\widehat{\pi}_{\mathrm{c}}}\big)^{-1} \widetilde{\boldsymbol{r}} - \big(\boldsymbol{I} - \gamma \widehat{\boldsymbol{P}}_{\widehat{\pi}_{\mathrm{c}}}\big)^{-1} \boldsymbol{r} = \big(\boldsymbol{I} - \gamma \widehat{\boldsymbol{P}}_{\widehat{\pi}_{\mathrm{c}}}\big)^{-1} \big(\widetilde{\boldsymbol{r}} - \boldsymbol{r}\big) \\ &\le \big\| \big(\boldsymbol{I} - \gamma \widehat{\boldsymbol{P}}_{\widehat{\pi}_{\mathrm{c}}}\big)^{-1} \big\|_1 \|\widetilde{\boldsymbol{r}} - \boldsymbol{r}\|_\infty \boldsymbol{1} \overset{\text{(ii)}}{=} \frac{1}{1-\gamma} \|\widetilde{\boldsymbol{r}} - \boldsymbol{r}\|_\infty \boldsymbol{1} \overset{\text{(iii)}}{\le} \frac{\varsigma}{1-\gamma} \boldsymbol{1} \overset{\text{(iv)}}{\le} \frac{\xi}{1-\gamma} \boldsymbol{1}. \end{aligned}$$

Here, (i) is due to the Bellman equation, (ii) relies on the fact $\big\| \big(\boldsymbol{I} - \gamma \widehat{\boldsymbol{P}}_{\widehat{\pi}_{\mathrm{c}}}\big)^{-1} \big\|_1 = \frac{1}{1-\gamma}$, (iii) arises since $\widehat{V}^\star(s) - \widehat{Q}^\star(s, \widehat{\pi}_{\mathrm{c}}(s)) \le \varsigma$ by construction of $\widehat{\pi}_{\mathrm{c}}$, whereas (iv) is valid since $\varsigma \in [0, \xi]$. The lemma then follows by recognizing that $\widehat{\boldsymbol{V}}^{\pi^\star} \le \widehat{\boldsymbol{V}}^\star$ due to the optimality of $\widehat{\boldsymbol{V}}^\star$.

EC.2.7. Proof of Lemma 8

We first make the key observation that, with probability at least $1 - \delta$, the following event holds true:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \qquad \big| \widehat{V}(s) - \widehat{Q}(s,a) - \varsigma \big| > \frac{\xi \delta}{2|\mathcal{S}||\mathcal{A}|}. \tag{EC.40}$$

To justify this claim, note that its complementary event satisfies

$$\mathbb{P}\left(\exists (s,a) \in \mathcal{S} \times \mathcal{A},\ |\widehat{V}(s) - \widehat{Q}(s,a) - \varsigma| \le \frac{\xi \delta}{2|\mathcal{S}||\mathcal{A}|} \right) \le \sum_{s,a} \mathbb{P}\left(\varsigma \in \left[\widehat{V}(s) - \widehat{Q}(s,a) \pm \frac{\xi \delta}{2|\mathcal{S}||\mathcal{A}|} \right] \right)$$

$$\leq |\mathcal{S}||\mathcal{A}| \cdot \frac{\xi\delta}{\xi|\mathcal{S}||\mathcal{A}|} = \delta,$$

where the first inequality applies the union bound, and the last line follows since $\varsigma \sim \mathsf{Unif}(0,\xi)$. Here, we abbreviate $[a \pm b] := [a-b, a+b]$. Combining (EC.40) with the assumption (55), we further reach

$$\left|V(s) - Q(s,a) - \varsigma\right| \geq \left|\widehat{V}(s) - \widehat{Q}(s,a) - \varsigma\right| - 2\max_{s,a}\left|Q(s,a) - \widehat{Q}(s,a)\right| > \frac{\xi\delta}{4|\mathcal{S}||\mathcal{A}|}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, or equivalently,

$$V(s) - Q(s,a) < \varsigma - \frac{\xi\delta}{4|\mathcal{S}||\mathcal{A}|} \quad \text{or} \quad V(s) - Q(s,a) > \varsigma + \frac{\xi\delta}{4|\mathcal{S}||\mathcal{A}|} \tag{EC.41}$$

holds for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

We are now prepared to justify the claim of this lemma. To begin with, consider any action $\widehat{a} \in \{a \in \mathcal{A} : Q(s,a) > V(s) - \varsigma\}$. Comparing this with the condition (EC.41), we can easily see that the only possibility is

$$V(s) - Q(s,\widehat{a}) < \varsigma - \frac{\xi\delta}{4|\mathcal{S}||\mathcal{A}|}.$$

Therefore, by invoking a basic decomposition, we ensure that

$$\begin{aligned}
\widehat{V}(s) - \widehat{Q}(s,\widehat{a}) &= \widehat{V}(s) - V(s) + V(s) - Q(s,\widehat{a}) + Q(s,\widehat{a}) - \widehat{Q}(s,\widehat{a}) \\
&\leq V(s) - Q(s,\widehat{a}) + \left|V(s) - \widehat{V}(s)\right| + \left|\widehat{Q}(s,\widehat{a}) - Q(s,\widehat{a})\right| \\
&< \varsigma - \frac{\xi\delta}{4|\mathcal{S}||\mathcal{A}|} + 2\max_{s,a}\left|Q(s,a) - \widehat{Q}(s,a)\right| \\
&\leq \varsigma - \frac{\xi\delta}{4|\mathcal{S}||\mathcal{A}|} + 2 \cdot \frac{\xi\delta}{8|\mathcal{S}||\mathcal{A}|} = \varsigma.
\end{aligned}$$

This essentially implies that

$$\left\{a \in \mathcal{A} : Q(s,a) > V(s) - \varsigma\right\} \subseteq \left\{a \in \mathcal{A} : \widehat{Q}(s,a) > \widehat{V}(s) - \varsigma\right\}. \tag{EC.42}$$

Applying exactly the same argument for any action $\widehat{a} \in \{a \in \mathcal{A} : Q(s,a) \leq V(s) - \varsigma\}$, we can also derive

$$\left\{a \in \mathcal{A} : \widehat{Q}(s,a) > \widehat{V}(s) - \varsigma\right\} \subseteq \left\{a \in \mathcal{A} : Q(s,a) > V(s) - \varsigma\right\}. \tag{EC.43}$$

These two set inequalities taken collectively establish the lemma.

EC.3. Proofs of auxiliary lemmas: finite-horizon MDPs

To begin, we find it helpful to control the entrywise magnitudes of $\{\boldsymbol{V}_h^{(l)}\}$. This is accomplished via the following lemma, with the proof postponed to Section EC.3.2.

LEMMA EC.3. *The vectors* $\left\{\boldsymbol{V}_h^{(l)}\right\}$ *defined in* (62) *obey*

$$\left\|\boldsymbol{V}_h^{(l)}\right\|_\infty \leq \left(\sqrt{3H}\right)^l H \tag{EC.44}$$

for all $1 \leq h \leq H$ *and all* $l \geq 0$.

EC.3.1. Proof of Lemma 9

Consider any l obeying $1 \le l \le m := \log_2 H$. By construction (cf. (62)), we see that

$$\begin{aligned}\widehat{\boldsymbol{V}}_h^{(l)} - \boldsymbol{V}_h^{(l)} &= \widehat{\boldsymbol{P}}_{h,\pi}\widehat{\boldsymbol{V}}_{h+1}^{(l)} - \boldsymbol{P}_{h,\pi}\boldsymbol{V}_{h+1}^{(l)} \\ &= \widehat{\boldsymbol{P}}_{h,\pi}\big(\widehat{\boldsymbol{V}}_{h+1}^{(l)} - \boldsymbol{V}_{h+1}^{(l)}\big) + \big(\widehat{\boldsymbol{P}}_{h,\pi} - \boldsymbol{P}_{h,\pi}\big)\boldsymbol{V}_{h+1}^{(l)},\end{aligned}$$

which connects $\boldsymbol{V}_h^{(l)}$ and $\widehat{\boldsymbol{V}}_h^{(l)}$ with $\boldsymbol{V}_{h+1}^{(l)}$ and $\widehat{\boldsymbol{V}}_{h+1}^{(l)}$. Apply the above relation recursively and make use of the conditions (61) to arrive at

$$\widehat{\boldsymbol{V}}_h^{(l)} - \boldsymbol{V}_h^{(l)} = \sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\widehat{\boldsymbol{P}}_{i,\pi}\big(\widehat{\boldsymbol{P}}_{j,\pi} - \boldsymbol{P}_{j,\pi}\big)\boldsymbol{V}_{j+1}^{(l)},$$

where we adopt the convenient notation and let $\prod_{i=h}^{h-1}\widehat{\boldsymbol{P}}_{i,\pi} = \boldsymbol{I}$.

According to the triangle inequality, we can further deduce that

$$\begin{aligned}\big|\widehat{\boldsymbol{V}}_h^{(l)} - \boldsymbol{V}_h^{(l)}\big| &\le \sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\widehat{\boldsymbol{P}}_{i,\pi}\Big|\big(\widehat{\boldsymbol{P}}_{j,\pi} - \boldsymbol{P}_{j,\pi}\big)\boldsymbol{V}_{j+1}^{(l)}\Big| \\ &\le \sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\widehat{\boldsymbol{P}}_{i,\pi}\left(\sqrt{\frac{\beta_1}{N}}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{j,\pi}}\big[\boldsymbol{V}_{j+1}^{(l)}\big]} + \frac{\beta_1\big\|\boldsymbol{V}_{j+1}^{(l)}\big\|_\infty}{N}\mathbf{1}\right) \\ &= \sqrt{\frac{\beta_1}{N}}\left(\sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\widehat{\boldsymbol{P}}_{i,\pi}\boldsymbol{r}_j^{(l+1)}\right) + \sum_{j=h}^{H-1}\frac{\beta_1\big\|\boldsymbol{V}_{j+1}^{(l)}\big\|_\infty}{N}\mathbf{1} \\ &\le \sqrt{\frac{\beta_1}{N}}\left(\sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\widehat{\boldsymbol{P}}_{i,\pi}\boldsymbol{r}_j^{(l+1)}\right) + \frac{\beta_1 H\max_j\big\|\boldsymbol{V}_j^{(l)}\big\|_\infty}{N}\mathbf{1},\end{aligned} \tag{EC.45}$$

where the second line follows from the assumption (63), and the third line makes use of the definition (62) of $\boldsymbol{r}_h^{(l)}$ and the elementary identity $\prod_{i=h}^{j-1}\widehat{\boldsymbol{P}}_{i,\pi}\mathbf{1} = \mathbf{1}$ (since each $\widehat{\boldsymbol{P}}_{i,\pi}$ is a probability transition matrix). In view of the construction (62), we can also derive recursively that

$$\widehat{\boldsymbol{V}}_h^{(l)} = \sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\widehat{\boldsymbol{P}}_{i,\pi}\boldsymbol{r}_j^{(l)}, \tag{EC.46}$$

which combined with (EC.45) yields

$$\big|\widehat{\boldsymbol{V}}_h^{(l)} - \boldsymbol{V}_h^{(l)}\big| \le \sqrt{\frac{\beta_1}{N}}\widehat{\boldsymbol{V}}_h^{(l+1)} + \frac{\beta_1 H\max_j\big\|\boldsymbol{V}_j^{(l)}\big\|_\infty}{N}\mathbf{1}. \tag{EC.47}$$

Further, the above inequality together with the triangle inequality immediately results in the following recursive relation

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}_h^{(l)} - \boldsymbol{V}_h^{(l)}\big\|_\infty &\le \sqrt{\frac{\beta_1}{N}}\big\|\widehat{\boldsymbol{V}}_h^{(l+1)}\big\|_\infty + \frac{\beta_1 H}{N}\max_j\big\|\boldsymbol{V}_j^{(l)}\big\|_\infty \\ &\le \sqrt{\frac{\beta_1}{N}}\big\|\widehat{\boldsymbol{V}}_h^{(l+1)} - \boldsymbol{V}_h^{(l+1)}\big\|_\infty + \sqrt{\frac{\beta_1}{N}}\big\|\boldsymbol{V}_h^{(l+1)}\big\|_\infty + \frac{\beta_1 H}{N}\max_j\big\|\boldsymbol{V}_j^{(l)}\big\|_\infty,\end{aligned}$$

thus revealing a useful connection between $\|\widehat{\boldsymbol{V}}_h^{(l)} - \boldsymbol{V}_h^{(l)}\|_\infty$ and $\|\widehat{\boldsymbol{V}}_h^{(l+1)} - \boldsymbol{V}_h^{(l+1)}\|_\infty$. Applying this relation recursively with a little algebra leads to

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}_h^{(0)} - \boldsymbol{V}_h^{(0)}\big\|_\infty \le \left(\sqrt{\frac{\beta_1}{N}}\right)^m \big\|\widehat{\boldsymbol{V}}_h^{(m)} - \boldsymbol{V}_h^{(m)}\big\|_\infty \\ + \sum_{l=1}^m \left(\sqrt{\frac{\beta_1}{N}}\right)^l \|\boldsymbol{V}_h^{(l)}\|_\infty + \frac{\beta_1 H}{N}\sum_{l=1}^m \left(\sqrt{\frac{\beta_1}{N}}\right)^{l-1} \max_j \big\|\boldsymbol{V}_j^{(l-1)}\big\|_\infty. \end{aligned} \tag{EC.48}$$

Additionally, it is easily seen from the definition (62) that

$$\boldsymbol{r}_h^{(l+1)} \le \boldsymbol{V}_{h+1}^{(l)} \le \big\|\boldsymbol{V}_{h+1}^{(l)}\big\|_\infty \mathbf{1},$$

which taken together with (EC.45) and the elementary identity $\prod_{i=h}^{j-1} \widehat{\boldsymbol{P}}_{i,\pi}\mathbf{1} = \mathbf{1}$ implies that

$$\begin{aligned}\big|\widehat{\boldsymbol{V}}_h^{(m)} - \boldsymbol{V}_h^{(m)}\big| &\le \sqrt{\frac{\beta_1}{N}} \max_j \big\|\boldsymbol{r}_j^{(m+1)}\big\|_\infty \left(\sum_{j=h}^{H-1}\prod_{i=h}^{j-1} \widehat{\boldsymbol{P}}_{i,\pi}\mathbf{1}\right) + \frac{\beta_1 H \max_j \big\|\boldsymbol{V}_j^{(m)}\big\|_\infty}{N}\mathbf{1} \\ &\le \sqrt{\frac{\beta_1}{N}} \max_j \big\|\boldsymbol{V}_j^{(m)}\big\|_\infty \left(\sum_{j=h}^{H-1}\mathbf{1}\right) + \frac{\beta_1 H \max_j \big\|\boldsymbol{V}_j^{(m)}\big\|_\infty}{N}\mathbf{1} \\ &\le H\left(\sqrt{\frac{\beta_1}{N}} + \frac{\beta_1}{N}\right) \max_j \big\|\boldsymbol{V}_j^{(m)}\big\|_\infty \mathbf{1}.\end{aligned}$$

Substitution into (EC.48) results in

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}_h^{(0)} - \boldsymbol{V}_h^{(0)}\big\|_\infty \le H\left(\sqrt{\frac{\beta_1}{N}}\right)^m \left(\sqrt{\frac{\beta_1}{N}} + \frac{\beta_1}{N}\right) \max_j \big\|\boldsymbol{V}_j^{(m)}\big\|_\infty \\ + \sum_{l=1}^m \left(\sqrt{\frac{\beta_1}{N}}\right)^l \|\boldsymbol{V}_h^{(l)}\|_\infty + \frac{\beta_1 H}{N}\sum_{l=1}^m \left(\sqrt{\frac{\beta_1}{N}}\right)^{l-1} \max_j \big\|\boldsymbol{V}_j^{(l-1)}\big\|_\infty. \end{aligned} \tag{EC.49}$$

To finish up, it remains to control the terms $\{\|\boldsymbol{V}_h^{(l)}\|_\infty\}$. Towards this, combining Lemma EC.3 with the above inequality (EC.49) yields

$$\begin{aligned}\big\|\widehat{\boldsymbol{V}}_h^{(0)} - \boldsymbol{V}_h^{(0)}\big\|_\infty &\le H^2\left(\sqrt{\frac{3\beta_1 H}{N}}\right)^m \left(\sqrt{\frac{\beta_1}{N}} + \frac{\beta_1}{N}\right) + \sum_{l=1}^m H\left(\sqrt{\frac{3\beta_1 H}{N}}\right)^l + \sum_{l=1}^m \frac{\beta_1 H^2}{N}\left(\sqrt{\frac{3\beta_1 H}{N}}\right)^{l-1} \\ &= H^2\left(\sqrt{\frac{3\beta_1 H}{N}}\right)^m \left(\sqrt{\frac{\beta_1}{N}} + \frac{\beta_1}{N}\right) + \left\{H\sqrt{\frac{3\beta_1 H}{N}} + \frac{\beta_1 H^2}{N}\right\}\sum_{l=1}^m \left(\sqrt{\frac{3\beta_1 H}{N}}\right)^{l-1} \\ &\le H^2\left(\sqrt{\frac{3\beta_1 H}{N}}\right)^m \left(\sqrt{\frac{\beta_1}{N}} + \frac{\beta_1}{N}\right) + 2\left\{H\sqrt{\frac{3\beta_1 H}{N}} + \frac{\beta_1 H^2}{N}\right\}, \end{aligned} \tag{EC.50}$$

Here, the last inequality holds true since

$$\sum_{l=1}^m \left(\sqrt{\frac{3\beta_1 H}{N}}\right)^{l-1} \le \frac{1}{1 - \sqrt{\frac{3\beta_1 H}{N}}} \le 2,$$

provided that $\sqrt{\frac{3\beta_1 H}{N}} \le 1/2$. Invoking the assumption $N \ge 12H\beta_1$ again and taking $m = \log_2 H$, we have $\big(\sqrt{\frac{3\beta_1 H}{N}}\big)^m \le 1/H$. This combined with (EC.51) immediately leads to

$$\big\|\widehat{\boldsymbol{V}}_h^{(0)} - \boldsymbol{V}_h^{(0)}\big\|_\infty \le 3\left\{H\sqrt{\frac{3\beta_1 H}{N}} + \frac{\beta_1 H^2}{N}\right\} \le 6H\sqrt{\frac{3\beta_1 H}{N}}, \tag{EC.51}$$

with the proviso that $N \ge 3\beta_1 H$. This concludes the proof.

EC.3.2. Proof of Lemma EC.3

To begin with, it is seen from the notation (28) that for all $1 \le j \le H$,

$$\begin{aligned}
\mathsf{Var}_{\boldsymbol{P}_{j,\pi}}\big(\boldsymbol{V}_{j+1}^{(l)}\big) &= \boldsymbol{P}_{j,\pi}\big(\boldsymbol{V}_{j+1}^{(l)} \circ \boldsymbol{V}_{j+1}^{(l)}\big) - \big(\boldsymbol{P}_{j,\pi}\boldsymbol{V}_{j+1}^{(l)}\big) \circ \big(\boldsymbol{P}_{j,\pi}\boldsymbol{V}_{j+1}^{(l)}\big) \\
&= \boldsymbol{P}_{j,\pi}\big(\boldsymbol{V}_{j+1}^{(l)} \circ \boldsymbol{V}_{j+1}^{(l)}\big) - \big(\boldsymbol{V}_{j}^{(l)} - \boldsymbol{r}_{j}^{(l)}\big) \circ \big(\boldsymbol{V}_{j}^{(l)} - \boldsymbol{r}_{j}^{(l)}\big) \\
&= \boldsymbol{P}_{j,\pi}\big(\boldsymbol{V}_{j+1}^{(l)} \circ \boldsymbol{V}_{j+1}^{(l)}\big) - \boldsymbol{V}_{j}^{(l)} \circ \boldsymbol{V}_{j}^{(l)} + 2\boldsymbol{V}_{j}^{(l)} \circ \boldsymbol{r}_{j}^{(l)} - \boldsymbol{r}_{j}^{(l)} \circ \boldsymbol{r}_{j}^{(l)} \\
&\le \boldsymbol{P}_{j,\pi}\big(\boldsymbol{V}_{j+1}^{(l)} \circ \boldsymbol{V}_{j+1}^{(l)}\big) - \boldsymbol{V}_{j}^{(l)} \circ \boldsymbol{V}_{j}^{(l)} + 2\boldsymbol{V}_{j}^{(l)} \circ \boldsymbol{r}_{j}^{(l)},
\end{aligned} \tag{EC.52}$$

where the second identity makes use of the fact that $\boldsymbol{V}_j^{(l)} = \boldsymbol{r}_j^{(l)} + \boldsymbol{P}_{j,\pi}\boldsymbol{V}_{j+1}^{(l)}$ (cf. (62)). Moreover, from the construction (62) we can easily derive

$$\boldsymbol{V}_h^{(l)} = \sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\boldsymbol{P}_{i,\pi}\boldsymbol{r}_j^{(l)} = \sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\boldsymbol{P}_{i,\pi}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{j,\pi}}\big[\boldsymbol{V}_{j+1}^{(l-1)}\big]}. \tag{EC.53}$$

The above two results taken collectively give

$$\begin{aligned}
\big|\boldsymbol{V}_h^{(l+1)}\big| &= \sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\boldsymbol{P}_{i,\pi}\sqrt{\mathsf{Var}_{\boldsymbol{P}_{j,\pi}}\big[\boldsymbol{V}_{j+1}^{(l)}\big]} \overset{\text{(i)}}{\le} \sum_{j=h}^{H-1}\sqrt{\prod_{i=h}^{j-1}\boldsymbol{P}_{i,\pi}\mathsf{Var}_{\boldsymbol{P}_{j,\pi}}\big[\boldsymbol{V}_{j+1}^{(l)}\big]} \\
&\overset{\text{(ii)}}{\le} \sqrt{H\sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\boldsymbol{P}_{i,\pi}\mathsf{Var}_{\boldsymbol{P}_{j,\pi}}\big[\boldsymbol{V}_{j+1}^{(l)}\big]} \\
&\overset{\text{(iii)}}{\le} \sqrt{H\sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\boldsymbol{P}_{i,\pi}\Big[\boldsymbol{P}_{j,\pi}\big(\boldsymbol{V}_{j+1}^{(l)} \circ \boldsymbol{V}_{j+1}^{(l)}\big) - \boldsymbol{V}_{j}^{(l)} \circ \boldsymbol{V}_{j}^{(l)} + 2\boldsymbol{V}_{j}^{(l)} \circ \boldsymbol{r}_{j}^{(l)}\Big]} \\
&= \sqrt{H\sum_{j=h}^{H-1}\Big(\prod_{i=h}^{j}\boldsymbol{P}_{i,\pi}\big(\boldsymbol{V}_{j+1}^{(l)} \circ \boldsymbol{V}_{j+1}^{(l)}\big) - \prod_{i=h}^{j-1}\boldsymbol{P}_{i,\pi}\big(\boldsymbol{V}_{j}^{(l)} \circ \boldsymbol{V}_{j}^{(l)}\big)\Big) + 2H\sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\boldsymbol{P}_{i,\pi}\big(\boldsymbol{V}_{j}^{(l)} \circ \boldsymbol{r}_{j}^{(l)}\big)}.
\end{aligned}$$

Here, (i) arises from Jensen's inequality; (ii) holds true due to the Cauchy-Schwarz inequality; (iii) follows from (EC.52). By telescoping summation, one further arrives at

$$\big|\boldsymbol{V}_h^{(l+1)}\big| \le \sqrt{H\Big[\prod_{i=h}^{H-1}\boldsymbol{P}_{i,\pi}\big(\boldsymbol{V}_{H}^{(l)} \circ \boldsymbol{V}_{H}^{(l)}\big) - \boldsymbol{V}_{h}^{(l)} \circ \boldsymbol{V}_{h}^{(l)}\Big] + 2H\max_j\big\|\boldsymbol{V}_j^{(l)}\big\|_\infty\sum_{j=h}^{H-1}\prod_{i=h}^{j-1}\boldsymbol{P}_{i,\pi}\boldsymbol{r}_j^{(l)}}$$

$$\overset{\text{(iv)}}{=} \sqrt{H\left[\prod_{i=h}^{H-1} \boldsymbol{P}_{i,\pi}\big(\boldsymbol{V}_H^{(l)} \circ \boldsymbol{V}_H^{(l)}\big) - \boldsymbol{V}_h^{(l)} \circ \boldsymbol{V}_h^{(l)}\right] + 2H \max_j \big\|\boldsymbol{V}_j^{(l)}\big\|_\infty \boldsymbol{V}_h^{(l)}}$$
$$\leq \sqrt{H \prod_{i=h}^{H-1} \boldsymbol{P}_{i,\pi}\big(\boldsymbol{V}_H^{(l)} \circ \boldsymbol{V}_H^{(l)}\big) + 2H\Big(\max_j \big\|\boldsymbol{V}_j^{(l)}\big\|_\infty\Big)^2 \mathbf{1}}$$
$$\overset{\text{(v)}}{\leq} \sqrt{3H} \max_j \big\|\boldsymbol{V}_j^{(l)}\big\|_\infty \mathbf{1}.$$

Here (iv) invokes the relation $\boldsymbol{V}_h^{(l)} = \sum_{j=h}^{H-1} \prod_{i=h}^{j-1} \boldsymbol{P}_{i,\pi} \boldsymbol{r}_j^{(l)}$ (see (EC.53)); and (v) holds true since

$$\left\| \prod_{i=h}^{H-1} \boldsymbol{P}_{i,\pi}\big(\boldsymbol{V}_H^{(l)} \circ \boldsymbol{V}_H^{(l)}\big) \right\|_\infty \leq \left\| \prod_{i=h}^{H-1} \boldsymbol{P}_{i,\pi} \right\|_1 \big\|\big(\boldsymbol{V}_H^{(l)} \circ \boldsymbol{V}_H^{(l)}\big)\big\|_\infty \leq \max_j \big\|\boldsymbol{V}_j^{(l)}\big\|_\infty^2.$$

As a consequence, the above inequality allows one to deduce that

$$\max_j \big\|\boldsymbol{V}_j^{(l+1)}\big\|_\infty \leq \sqrt{3H} \max_j \big\|\boldsymbol{V}_j^{(l)}\big\|_\infty,$$

and therefore,

$$\max_j \big\|\boldsymbol{V}_j^{(l)}\big\|_\infty \leq \big(\sqrt{3H}\big)^l \max_j \big\|\boldsymbol{V}_j^{(0)}\big\|_\infty \leq \big(\sqrt{3H}\big)^l H,$$

where the last inequality arises from the trivial upper bound $\max_j \|\boldsymbol{V}_j^{(0)}\|_\infty \leq H$.

Endnotes

1. Here and throughout, the "model" refers to the transition kernel and the rewards of the MDP taken collectively.

2. Note that perturbation is only invoked when running the planning algorithms and does not require collecting new samples.

Acknowledgments

Y. Wei is supported in part by the the NSF grants CCF-2106778, DMS-2147546/2015447 and CAREER award DMS-2143215. Y. Chi is supported in part by the ONR grants N00014-18-1-2142 and N00014-19-1-2404, the NSF grants CCF-1806154, CCF-2007911 and CCF-2106778. Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grants FA9550-19-1-0030 and FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009, CCF-1907661, DMS-2014279, IIS-2218713 and IIS-2218773. We thank Qiwen Cui for pointing out an issue in Section EC.2.5 in an early version of this paper, and thank Shicong Cen, Chen Cheng and Cong Ma for numerous discussions. Part of this work was done while G. Li, Y. Wei and Y. Chen were visiting the Simons Institute for the Theory of Computing.

References

Agarwal A, Kakade S, Yang LF (2020) Model-based reinforcement learning with a generative model is minimax optimal. *Conference on Learning Theory (COLT)* .

Azar MG, Munos R, Kappen HJ (2013) Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning* 91(3):325–349.