A Proofs

Lemma 15 (A one-step good policy is close to optimal)**.** *Let* $\Delta(h) := |V^*_\xi(h) - V^\pi_\xi(h)|$ *with* $h \in (\mathcal{A}\times\mathcal{E})^t$ *for* $t \geq t_0 \in \mathbb{N}$.

$$\begin{aligned} \textit{If} \quad & \mathbb{E}^\pi_\xi |\max_a Q^\pi_\xi(h,a) - V^\pi_\xi(h)| < \beta \quad \forall t \geq t_0 \\ \textit{and} \quad & \mathbb{E}^\pi_\xi [\max_a \textstyle\sum_e \xi(e|ha)\Delta(hae)] \leq (1+\alpha)\mathbb{E}^\pi_\xi \Delta(hae) \quad \forall t \geq t_0 \\ \textit{then} \quad & \mathbb{E}^\pi_\xi \Delta(h) < \tfrac{\beta}{1-\gamma(1+\alpha)} \quad \forall t \geq t_0 \quad \textit{provided} \quad 1+\alpha < 1/\gamma \end{aligned}$$

Proof. Let $\delta := \sup_{t\geq t_0} \mathbb{E}\Delta(h)$, where $h \in (\mathcal{A}\times\mathcal{E})^t$ and $\mathbb{E}$ is short for $\mathbb{E}^\pi_\xi$.

$$\begin{aligned} \mathbb{E}\Delta(h) &= \big|\max_a Q^*_\xi(h,a) - V^\pi_\xi(h)\big| \\ &= \mathbb{E}\big|\max_a Q^*_\xi(h,a) - \max_a Q^\pi_\xi(h,a) + \max_a Q^\pi_\xi(h,a) - V^\pi_\xi(h)\big| \\ &\leq \mathbb{E}\big|\max_a Q^\pi_\xi(h,a) - V^\pi_\xi(h)\big| + \mathbb{E}\big|\max_a Q^*_\xi(h,a) - \max_a Q^\pi_\xi(h,a)\big| \\ &\overset{(1)}{<} \beta + \mathbb{E}\Big|\max_a \sum_e \xi(e|ha)\left(r + \gamma V^*_\xi(hae)\right) - \max_a \sum_e \xi(e|ha)\left(r+\gamma V^\pi_\xi(hae)\right)\Big| \\ &\leq \beta + \gamma\mathbb{E}\max_a \sum_e \xi(e|ha)|V^*_\xi(hae) - V^\pi_\xi(hae)| \\ &\leq \beta + \gamma(1+\alpha)\mathbb{E}\Delta(hae) \end{aligned}$$

Taking $\sup_{t\geq t_0}$ on both sides implies $\delta < \beta + \gamma(1+\alpha)\delta$ implies $\delta < \beta/(1-\gamma(1+\alpha))$. □ □

Lemma 17 ($\mathbb{E}^\pi_\xi \to 0$ implies $\mathbb{E}^\pi_\mu \to 0$)**.** *If* π *is such that*

$$\mathbb{E}^\pi_\xi\left[V^*_\xi(h_{<t}) - V^\pi_\xi(h_{<t})\right] \to 0 \quad \textit{as} \quad t\to\infty.$$

then for all $\mu \in \mathcal{M}$ *we have*

$$\mathbb{E}^\pi_\mu\left[V^*_\xi(h_{<t}) - V^\pi_\xi(h_{<t})\right] \to 0 \quad \textit{as} \quad t\to\infty.$$

Proof.

$$\mathbb{E}^\pi_\mu\left[V^*_\xi(h_{<t}) - V^\pi_\xi(h_{<t})\right] \leq \frac{1}{w(\mu)}\mathbb{E}^\pi_\xi\left[V^*_\xi(h_{<t}) - V^\pi_\xi(h_{<t})\right] \to 0$$

by the dominance of $\xi(\cdot) \geq w(\mu)\mu(\cdot)$. □

Lemma 20 ($V^{\pi'}_\xi \to V^\pi_\xi$ implies $V^{\pi'}_\mu \to V^\pi_\mu$ in μ-expectation)**.** *If* π *is such that for all* $\mu \in \mathcal{M}$

$$\mathbb{E}^\pi_\mu\left[V^{\pi'}_\xi(h_{<t}) - V^\pi_\xi(h_{<t})\right] \to 0 \quad \textit{as} \quad t\to\infty.$$

and $D_\infty(\mu^{\pi'}, \xi^{\pi'}|h_{<t}) \to 0$ μ^π*-almost surely then we have*

$$\mathbb{E}^\pi_\mu\left[V^{\pi'}_\mu(h_{<t}) - V^\pi_\mu(h_{<t})\right] \to 0 \quad \textit{as} \quad t\to\infty.$$

Proof.

$$\begin{aligned} &\mathbb{E}^\pi_\mu\left[|V^{\pi'}_\mu(h_{<t}) - V^\pi_\mu(h_{<t})|\right] \\ &= \mathbb{E}^\pi_\mu\left[|V^{\pi'}_\mu(h_{<t}) - V^{\pi'}_\xi(h_{<t}) + V^{\pi'}_\xi(h_{<t}) - V^\pi_\xi(h_{<t}) + V^\pi_\xi(h_{<t}) - V^\pi_\mu(h_{<t})|\right] \\ &\leq \mathbb{E}^\pi_\mu\left[|V^{\pi'}_\mu(h_{<t}) - V^{\pi'}_\xi(h_{<t})|\right] + \mathbb{E}^\pi_\mu\left[|V^{\pi'}_\xi(h_{<t}) - V^\pi_\xi(h_{<t})|\right] + \mathbb{E}^\pi_\mu\left[|V^\pi_\xi(h_{<t}) - V^\pi_\mu(h_{<t})|\right] \end{aligned}$$

The second and third term go to 0 as $t \to \infty$ by the assumptions and Lemma 3 with Lemma 13. The first term goes to 0 as $D_\infty(\mu^{\pi'}, \xi^{\pi'} | h_{<t}) \to 0$ μ^π-almost surely implies $\mathbb{E}^\pi_\mu \left[D_\infty(\mu^{\pi'}, \xi^{\pi'} | h_{<t}) \right] \to 0$ and we have $\mathbb{E}^\pi_\mu \left[|V^{\pi'}_\mu(h_{<t}) - V^{\pi'}_\xi(h_{<t})| \right] \le \mathbb{E}^\pi_\mu \left[D_\infty(\mu^{\pi'}, \xi^{\pi'} | h_{<t}) \right]$.

□

Theorem 22 (Self-AIXI is Self-optimizing)**.** *Let μ be some environment. If there is a policy π and a sequence of policies $\overline{\pi_1}, \overline{\pi_2} \dots$ all contained within $\mathcal{P}$ such that for all $t, h_{<t}$ we have $V^\zeta_\xi(h_{<t}) \ge V^{\overline{\pi_t}}_\xi(h_{<t}) - \epsilon_t$ with $\epsilon_t \to 0$, and for all $\nu \in \mathcal{M}$*

$$V^*_\nu(h_{<t}) - V^{\overline{\pi_t}}_\nu(h_{<t}) \to 0 \quad \textit{as} \quad t \to \infty \quad \mu^\pi\textit{-almost surely} \tag{4}$$

then

$$V^*_\nu(h_{<t}) - V^{\pi_S}_\nu(h_{<t}) \to 0 \quad \textit{as} \quad t \to \infty \quad \mu^\pi\textit{-almost surely}$$

If $\pi = \pi_S$ and Equation 4 holds for all $\mu \in \mathcal{M}$, then π_S is strongly asymptotically optimal in the class $\mathcal{M}$.

Proof.

$$\begin{aligned}
0 &\le w(\mu|h_{<t}) \left(V^*_\mu(h_{<t}) - V^{\pi_S}_\mu(h_{<t}) \right) & (5)\\
&\le \sum_{\nu \in \mathcal{M}} w(\nu|h_{<t}) \left(V^*_\nu(h_{<t}) - V^{\pi_S}_\nu(h_{<t}) \right) & (6)\\
&= \sum_{\nu \in \mathcal{M}} w(\nu|h_{<t}) V^*_\nu(h_{<t}) - V^{\pi_S}_\xi(h_{<t}) & (7)\\
&\le \sum_{\nu \in \mathcal{M}} w(\nu|h_{<t}) V^*_\nu(h_{<t}) - V^\zeta_\xi(h_{<t}) & (8)\\
&\le \sum_{\nu \in \mathcal{M}} w(\nu|h_{<t}) V^*_\nu(h_{<t}) - V^{\overline{\pi_t}}_\xi(h_{<t}) + \epsilon_t & (9)\\
&= \sum_{\nu \in \mathcal{M}} w(\nu|h_{<t}) \left(V^*_\nu(h_{<t}) - V^{\overline{\pi_t}}_\nu(h_{<t}) \right) + \epsilon_t & (10)\\
&\to 0 & (11)
\end{aligned}$$

Equation 6 comes from adding positive terms. Equations 7 and 10 comes from the linearity of the value function. Equation 8 comes from π_S being one step optimal then following ζ and . Equation 9 comes from the assumptions. Lastly, 11 comes from Equation 4 and [14, Lem.5.28ii].

$w(\mu|h_{<t}) \not\to 0$ as $h_{<t}$ is generated from μ^π (for more details see Self-Optimizing proof in [14]). Therefore $V^*_\mu(h_{<t}) - V^{\pi_S}_\mu(h_{<t}) \to 0$ μ^π-almost surely.

□