Optimality of solutions via LMDPs

> Do these two paths lead to the same place?

One of the main questions we have not addressed yet is; if we solve the MDP directly, or solve it via our linear abstraction (linearise, solve and project), do we end up in the same place? This is a question about the completeness of our abstraction. Can our abstraction represent (and find) the same solutions that the original can?

Why does this matter? When we apply our linear abstraction, we want to know: can I trust the answer it has given? If I follow the actions specified by optimal policy, am I going to get rewards?

$$
\begin{aligned}
\| V_{\pi^*} - V_{\pi_{u^*}} \|_\infty &= \epsilon && (1)\\
&= \| (I - \gamma P_{\pi^*})^{-1} r_{\pi^*} - (I - \gamma P_{\pi_{u^*}})^{-1} r_{\pi_{u^*}} \|_\infty && (2)\\
&\leq \| (I - \gamma P_{\pi^*})^{-1} r_{\text{max}} - (I - \gamma P_{\pi_{u^*}})^{-1} r_{\text{min}} \|_\infty && (3)\\
&= \| \left((I - \gamma P_{\pi^*})^{-1} - (I - \gamma P_{\pi_{u^*}})^{-1} \right) \Delta r \|_\infty && (4)\\
&\leq \Delta r_{\text{max}} \| (I - \gamma P_{\pi^*})^{-1} - (I - \gamma P_{\pi_{u^*}})^{-1} \|_\infty && (5)\\
&= \Delta r_{\text{max}} \| \sum_{t=0}^{\infty} \gamma^t P_{\pi^*} - \sum_{t=0}^{\infty} \gamma^t P_{\pi_{u^*}} \|_\infty && (6)\\
&= \Delta r_{\text{max}} \| \sum_{t=0}^{\infty} \gamma^t (P_{\pi^*} - P_{\pi_{u^*}}) \|_\infty && (7)\\
&= \frac{\Delta r_{\text{max}}}{1 - \gamma} \| P_{\pi^*} - P_{\pi_{u^*}} \|_\infty && (8)
\end{aligned}
$$

(1)

(1) We want to compare the value achieved by the optimal policy and the value achieved by the optimal linearised solution.

(2) Assume that there exists a policy that can generate the optimal control dynamics (as given by the LMDP). In that case we can set $P_{\pi_{u^*}} = U^*$.

(3) r_{u^*} doesnt really make sense as the reward is action dependent. We could calculate it as $r_{\pi_{u^*}}$, but we dont explicity know π_{u^*}. $(I - \gamma P_{\pi^*})^{-1} r$ represents the action-values, or Q values. By doing this exhange, we might over estimate the diffference under the infinity norm as two non-optimal actions may have larger difference. Also, use the element wise infinity norm.

(4) Let's assume the optimal policy picks $\max_a r(s,a)$ at every step. Then the worst case is that we pick $\min_a r(s,a)$. Write $\max_a r(s,a) - \min_a r(s,a) = \Delta r$

(5) No. Cannot do that… Assumes $(I - \gamma P_{\pi^*})^{-1} = (I - \gamma P_{\pi_{u^*}})$. Or that the eqn can be factored.

Notes

- why are we using the infinity norm?!!
- What does $\delta \geq \mathrm{KL}(U^* \parallel P_{\pi_{u^*}})$ imply about $\parallel P_{\pi^*} - P_{\pi_{u^*}} \parallel_\infty$?
- Is it possible to relate U^* to π^*? $\mathrm{KL}(U^* \parallel P_{\pi^*})$
- Use $r \in [0, 1]$. Like in other papers. This would simplify and allow us to do (4)?? But need to check it generalises.

Alternative.

$$\parallel V_{\pi^*} - V_{\pi_{u^*}} \parallel_\infty = \epsilon \tag{1}$$

$$= \parallel \max_a \left[r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} [V(s')] \right] - \mathop{\mathbb{E}}_{a \sim \pi_{u^*}(\cdot|s)} r(s,a) - \gamma \mathop{\mathbb{E}}_{s' \sim \sum_a P(\cdot|s,a)\pi_{u^*}(a|s)} [V(s')] \parallel_\infty \tag{2}$$

$$= \parallel \max_a \left[r(s,a) - \mathop{\mathbb{E}}_{\tilde{a} \sim \pi_{u^*}(\cdot|s)} r(s,\tilde{a}) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} [V(s')] - \gamma \mathop{\mathbb{E}}_{s' \sim \sum_{\tilde{a}} P(\cdot|s,\tilde{a})\pi_{u^*}(\tilde{a}|s)} [V(s')] \right] \parallel_\infty$$

(2)

(2)

Relies a lot on the result of the projection from optimal state distributions to state-action policies.

$$P_\pi(\cdot|s) = \sum_a P(\cdot|s,a)\pi(a|s) \tag{3}$$

$$\pi = \mathop{\mathrm{argmin}}_\pi \mathrm{KL}\Big(u(\cdot|s)) \parallel P_\pi(\cdot|s) \Big) \tag{4}$$

$$\delta = \mathrm{KL}\Big(u^*(\cdot|s)) \parallel P_\pi(\cdot|s) \Big) \tag{5}$$

$$\delta = -\sum_{s'} u^*(s'|s) \log \frac{P_\pi(s'|s)}{u^*(s'|s)} \tag{6}$$

Small delta implies ???. Everywhere u^* has non-zero probability, $P_\pi(s'|s)$, is $\approx u(s'|s)$

KL

> the expected number of extra bits required to code samples from P using a code optimized for Q rather than the code optimized for P

Always gives more weight to the high probability (under u) states. And because the two distributions

are normalised. For all mass that $P_\pi > u$, there will be equal mass where $P_\pi < u$. But the latter is weighted more. (kinda)

---

want to know the difference between the two policies.

$$\delta = \mathsf{KL}\Big(\pi^* \parallel \pi_{u^*}\Big) \tag{7}$$

(8)

Could then use this to relate r_{π^*} and $r_{\pi_{u^*}}$? But how!?!