Maximisation derivation

Pick $a \in A$, versus, pick $\Delta(S)$. $f : S \to A$ vs $f : S \to \Delta(S)$.

In the original Todorov paper, they derive the LMDP equations for minimising a cost function. This maximisation derivation just changes a few negative signs around. Although there is also a change in the interpretation of what the unconstrained dynamics are doing. ...?

$$V(s) = \max_u q(s) - \mathsf{KL}(u(\cdot|s) \parallel p(\cdot|s)) + \gamma \mathop{\mathbb{E}}_{s' \sim u(\cdot|s)} V(s') \tag{1}$$

(1)

$$= q(s) + \max_u \left[\mathop{\mathbb{E}}_{s' \sim u(\cdot|s)} \log(\frac{p(s'|s)}{u(s'|s)} + \gamma \mathop{\mathbb{E}}_{s' \sim u(\cdot|s)} [V(s')] \right] \tag{2}$$

$$\log(z(s)) = q(s) + \max_u \left[\mathop{\mathbb{E}}_{s' \sim u(\cdot|s)} \log(\frac{p(s'|s)}{u(s'|s)} + \mathop{\mathbb{E}}_{s' \sim u(\cdot|s)} [\log(z(s')^\gamma)] \right] \tag{3}$$

$$= q(s) + \max_u \left[\mathop{\mathbb{E}}_{s' \sim u(\cdot|s)} \log(\frac{p(s'|s) z(s')^\gamma}{u(s'|s)}) \right] \tag{4}$$

$$G(s) = \sum_{s'} p(s'|s) z(s')^\gamma \tag{5}$$

$$= q(s) + \max_u \left[\mathop{\mathbb{E}}_{s' \sim u(\cdot|s)} \log(\frac{p(s'|s) z(s')^\gamma}{u(s'|s)} \cdot \frac{G(s)}{G(s)}) \right] \tag{6}$$

$$= q(s) + \log G(s) + \min_u \left[\mathsf{KL}(u(\cdot|s) \parallel \frac{p(\cdot|s) \cdot z(\cdot)^\gamma}{G(s)}) \right] \tag{7}$$

$$u^*(\cdot|s) = \frac{p(\cdot|s) \cdot z(\cdot)^\gamma}{\sum_{s'} p(s'|s) z(s')^\gamma} \tag{8}$$

$$\log(z_{u^*}(s)) = q(s) + \log\big(\sum_{s'} p(s'|s) z_{u^*}(s')^\gamma\big) \tag{9}$$

$$z_{u^*}(s) = e^{q(s)}\big(\sum_{s'} p(s'|s) z_{u^*}(s')^\gamma\big) \tag{10}$$

$$z_{u^*} = e^{q(s)} \cdot P z_{u^*}^\gamma \tag{11}$$

(2)

By definition, an LMDP is the optimisation problem in (1). We can move the max in (2), as $q(s)$ is not a function of u. Also in (2), expand the second term using the definiton of KL divergence, the negative from the KL cancels the second terms negative. (3) Define a new variable, $z(s) = e^{v(s)}$. Also, use the log rule to move the discount rate. (4) Both expectations are under the same distribution, therefore they can be combined. Also, using log rules, combine the log terms. (5) Define a new variable that will be used to normalise $p(s'|s)z(s')^\gamma$. (6) Multiply and divide by $G(s)$. This allows us to rewrite the log term as a KL divergence as now we have two distributions, $u(\cdot|s)$ and $\frac{p(\cdot|s)z(\cdot)^\gamma}{G(s)}$. (7) The change to a KL term introduces a negative, instead of maximising the negative KL, we minimise the KL. Also in (7) the extra G(s) term can be moved outside of the expectation as it is not dependent in s'. (8) Set the optimal policy to minimise the KL distance term. (9) Since we picked the optimal control to be the form in (8), the KL divergence term is zero. (10) Move the log. (11) Rewrite the equations for the tabular setting, giving a z vector, uncontrolled dynamics matrix.

MDP Linearisation

Goal. Take a MDP and find a LMDP that has similar structure.

NOTE: The derivation is not the same as in Todorov. He sets $b_a \neq r, b_a = r - \sum P \log P$.

$$\begin{aligned}
\forall s, s' \in S, \forall a \in A, \exists u_a \ \text{ such that;} \quad & \tag{1} \\
P(s'|s,a) &= u_a(s'|s)p(s'|s) \tag{2} \\
r(s,a) &= q(s) - \text{KL}(P(\cdot|s,a) \parallel u_a(\cdot|s)) \tag{3}
\end{aligned}$$

$$\tag{3}$$

$$\begin{aligned}
r(s,a) &= q(s) - \text{KL}(P(\cdot|s,a) \parallel \frac{P(\cdot|s,a)}{p(\cdot|s)}) \tag{4} \\
r(s,a) &= q(s) - \sum_{s'} P(s'|s,a) \log(p(s'|s)) \tag{5}
\end{aligned}$$

$$\tag{4}$$

$$\begin{aligned}
m_{s'}[s] &:= \log p(s'|s) \tag{6} \\
D_{as'}[s] &:= p(s'|s,a) \tag{7} \\
c_{s'}[s] &:= q[s]\mathbf{1} - m_{s'}[s] \ \text{ such that } \sum_{s'} e^{m_{s'}[s]} = 1 \tag{8}
\end{aligned}$$

$$\tag{5}$$

$$\begin{aligned}
r_a &= D_{as'}(q\mathbf{1} - m_{s'}) \ \ \forall s \tag{9} \\
r_a &= D_{as'}c_{s'} \ \ \forall s \tag{10} \\
c_{s'} &= r_a D^{\dagger}_{as'} \ \ \forall s \tag{11} \\
q &= -\log \sum_{s'} e^{-c_{s'}} \ \ \forall s \tag{12} \\
m_{s'} &= q - c_{s'} \ \ \forall s \tag{14}
\end{aligned}$$

$$\tag{6}$$

Therefore we have $|A|$ linear relationships, for each $s \in S$.

(5) KL introduces a negative, but we also use the log rule to move $p(s'|s)$ to the nominator, adding another negative, and thus cancelling. Yielding the cross entropy between .. and … .

(6) The More-Penrose pseudo inverse. We apply to the RHS… ???.

> If D is row-rank defficient, the solution c is not unique, and we should be able to exploit the freedom in choosing c to improve the approximation of the traditional MDP. If D is coumn-rank-defficient then an exact embedding cannot be constructed. However, this is unlikely in to occur in practice because it essentially means that the number of symbolic actions is greater than the number of possible next states.