Report

June 10, 2022

1 Entropy Regularized POMDP Model

Assume when $\alpha = 1$, namely using standard Gumbel distribution, the related functions and parameters are $U_\tau^{(1)}, V_t^{(1)}, Q^{(1)}, r^{(1)}, \pi^{(1)}, \theta^{(1)}$. Then for any $\alpha \neq 1$, then we have the following equations

- $U_\tau^\alpha = \alpha \cdot U_\tau^{(1)}$
- $V_t^\alpha = \alpha \cdot V_t^{(1)}$
- $Q_t^\alpha = \alpha \cdot Q_t^{(1)}$
- $r^\alpha = \alpha \cdot r^{(1)}$
- $\theta_1^\alpha = \alpha \cdot \theta_1^{(1)}$
- $\pi^\alpha = \pi^{(1)}$
- $\sigma^\alpha = \sigma^{(1)}$
- $\lambda^\alpha = \lambda^{(1)}$
- $\theta_2^\alpha = \theta_2^{(1)}$

Parameter	α	$\theta_{3,0,0}$	$\theta_{3,0,1}$	$\theta_{3,0,2}$	$\theta_{3,1,0}$	$\theta_{3,1,1}$	$\theta_{3,1,2}$	$\theta_{2,0}$	$\theta_{2,1}$	$\theta_{1,0}$	$\theta_{1,1}$	RC
Good State	1	0.039	0.333	0.590	$*$	$*$	$*$	0.949	$*$	0.219	$*$	9.257
	0.1									0.022		0.926
	0.01									0.002		0.093
Bad State	1	$*$	$*$	$*$	0.181	0.759	0.061	$*$	0.988	$*$	1.165	
	0.1										0.116	
	0.01										0.012	
log-Likelihood	-3819											

TABLE 1. PARAMETER ESTIMATES AND LOG-LIKELIHOOD PROVIDED BY THE POMDP MODEL ON GROUP 4 DATA SET

2 Appendix: Reasons

$$
\begin{aligned}
U_{\tau,\theta}(h_\tau) &= \sup_{\pi\in\Pi} \mathbb{E}\Bigg[\sum_{t\geq\tau} \beta^{t-\tau}[r_{\theta_1}(z_t, s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|h_t))]\Bigg] \\
&= \sup_{\pi\in\Pi} \mathbb{E}\Bigg[r_{\theta_1}(z_\tau, s_\tau, a_\tau) + \alpha\mathcal{H}(\pi(\cdot|h_\tau)) \\
&+\beta \sum_{t\geq\tau+1} \beta^{t-(\tau+1)}[r_{\theta_1}(z_t, s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|h_t))]\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
U_{t,\theta}(h_t) = \max_{\pi(\cdot|h_t)} \Big\{ &\sum_{a_t}\sum_{s_t} r_{\theta_1}(z_t, s_t, a_t)x_{t,\theta_2}(s_t)\pi(a_t|h_t) \\
&+ \alpha\mathcal{H}(\pi(\cdot|h_t)) \\
&+ \beta \sum_{z_{t+1}}\sum_{a_t} \mathbb{P}_{\theta_2}(z_{t+1}|h_t, a_t)\pi(a_t|h_t)U_{t+1,\theta}(h_{t+1})\Big\}.
\end{aligned} \tag{2.1}
$$

$$
\sigma_{\theta_2}(z_{t+1}, z_t, x_{t,\theta_2}, a_t) \triangleq \sum_{s'}\sum_{s} x_{t,\theta_2}(s)\mathbb{P}_{\theta_2}(z_{t+1}, s'|z_t, s, a_t),
$$

$$
\lambda_{\theta_2}(z_{t+1}, z_t, x_{t,\theta_2}, a_t) \triangleq \frac{x_{t,\theta_2}^\top P_{\theta_2}(z_{t+1}, z_t, a_t)}{\sigma_{\theta_2}(z_{t+1}, z_t, x_{t,\theta_2}, a_t)}, \tag{2.2}
$$

assuming $\sigma_{\theta_2}(z_{t+1}, z_t, x_{t,\theta_2}, a_t) \neq 0$, where we denote the (s, s') element of the matrix $[P_{\theta_2}(z_{t+1}, z_t, a_t)]_{s,s'} \triangleq \mathbb{P}_{\theta_2}(z_{t+1}, s'|z_t, s, a_t)$, $s, s' \in S$, $x_{t,\theta_2}^\top$ is the transpose of x_{t,θ_2}, and

$$
\begin{aligned}
&[x_{t,\theta_2}^\top P_{\theta_2}(z_{t+1}, z_t, a_t)]_{s_{t+1}} \\
&\qquad\triangleq \sum_{s} x_{t,\theta_2}(s)\mathbb{P}_{\theta_2}(z_{t+1}, s_{t+1}|z_t, s, a_t).
\end{aligned}
$$

$$
r_{\theta_1}(z_t, x_{t,\theta_2}, a_t) \triangleq \sum_{s} r_{\theta_1}(z_t, s, a_t)x_{t,\theta_2}(s).
$$

Proposition 1.

$$
\begin{aligned}
V_{t,\theta}(z, x) = \max_{\pi(\cdot|z,x)} \Big\{ &\sum_{a} r_{\theta_1}(z, x, a)\pi(a|z, x) + \alpha\mathcal{H}(\pi(\cdot|z, x)) \\
&+ \beta \sum_{a}\sum_{z'} \sigma_{\theta_2}(z', z, x, a)V_{t+1,\theta}(z', x'(a))\pi(a|z, x)\Big\}
\end{aligned}
$$

$$
[\mathcal{B}_\theta Q](z, x, a) = r_{\theta_1}(z, x, a) + \beta \sum_{z'} \sigma_{\theta_2}(z', z, x, a)V(z', x'), \tag{2.3}
$$

where $x' = \lambda_{\theta_2}(z', z, x, a)$ and

$$
\frac{1}{\alpha}V(z, x) \triangleq \max_{\hat{\pi}(\cdot|z,x)} \Big[\sum_{a} \frac{1}{\alpha}Q(z, x, a)\hat{\pi}(a|z, x) + \mathcal{H}(\hat{\pi}(\cdot|z, x))\Big].
$$

Theorem 1. *(a) $V(z, x) = \alpha \log\sum_a \exp(\frac{1}{\alpha}Q(z, x, a))$ for all $Q \in \mathcal{Q}$. (b) $\mathcal{B}_\theta : \mathcal{Q} \to \mathcal{Q}$ is a contraction mapping with modulus $\beta \in (0, 1)$ and (c) the optimal policy is of the form:*

$$
\pi_\theta(a|z, x) = \frac{\exp \frac{1}{\alpha}Q_\theta(z, x, a)}{\sum_{a'\in A} \exp \frac{1}{\alpha}Q_\theta(z, x, a')}, \tag{2.4}
$$

where Q_θ is the unique fixed point of $\mathcal{B}_\theta$ (i.e. $Q_\theta = \mathcal{B}_\theta Q_\theta$).