A Theory of Regularized MDPs

Peng Lingwei

August 29, 2019

Contents

1 Regularized MDPs

1. Regularized function: $\Omega(\pi)$ is strongly convex;

2. Regularized value functions: $V^{\pi,\Omega}(s) = V^{\pi} - \Omega(\pi(s))$

$$\begin{aligned}
V^{\pi,\Omega}(s) =& \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t(r(S_t,A_t)-(1-\lambda)\Omega(\pi(s)))|S_0=s\right]\\
=& \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t(r(S_t,A_t))|S_0=s\right]-\sum_{t=0}^{\infty}(1-\gamma)\Omega(\pi(s))\\
=& V^{\pi}(s)-\Omega(\pi(s))
\end{aligned}$$

In MDP, $Q^{\pi}(s,a) = r(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}\left[V^{\pi}(s')\right]$. And $V^{\pi} = T^{\pi}V^{\pi} = (\langle\pi(s), Q^{\pi}(s,\cdot)\rangle)_{s\in S}$. Then,let $Q^{\pi,\Omega}(s,a) = r(s,a)+\gamma\mathbb{E}_{P(s'|s,a)}\left[V^{\pi,\Omega}(s')\right]$,

$$V^{\pi,\Omega}(s) = \langle\pi(s), Q^{\pi,\Omega}(s,\cdot)\rangle - (1-\gamma)\Omega(\pi(s))$$

3. Regularized optimal value function: $V^{*,\Omega}(s) = \max_{\pi\in\Pi^{MR}} V^{\pi}(s)-\Omega(\pi(s))$
Let $Q^{*,\Omega}(s,\cdot) = r(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}\left[V^{*,\Omega}(s')\right]$.

$$\begin{aligned}
V^{*,\Omega}(s) =& \max_{\pi\in\Pi^{MR}} V^{\pi}(s)-\Omega(\pi(s))\\
=& \max_{\pi\in\Pi^{MR}}\langle\pi(s), Q^{\pi,\Omega}(s,\cdot)\rangle-(1-\gamma)\Omega(\pi(s))\\
=& \max_{\pi\in\Pi^{MR}}\langle\pi(s), Q^{*,\Omega}(s,\cdot)\rangle-(1-\gamma)\Omega(\pi(s)) \quad \text{(proof is trivial)}\\
=& \Omega^{*}_{\gamma}(Q^{*,\Omega}(s,\cdot))
\end{aligned}$$

where Ω^*_γ is Legendre-Fenchel transform of $(1-\gamma)\Omega$. More specifically,

$$\forall q_s\in\mathbb{R}^{|A|}, \Omega^*_\gamma(q_s) = \max_{\pi\in\Pi^{MR}}\langle\pi_s, q_s\rangle-(1-\gamma)\Omega(\pi_s)$$

4. Regularized Bellman operator: $T^{\pi,\Omega}V = T^{\pi}V-(1-\gamma)\Omega(\pi)$

   - Let $Q_V(s,a) = r(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}\left[V(s')\right]$,

   $$T^{\pi,\Omega}V(s) = \langle\pi_s, Q_V(s,\cdot)\rangle-(1-\lambda)\Omega(\pi_s)$$

   - Monotonicity:$V_1\succeq V_2 \Rightarrow T^{\pi,\Omega}V_1\succeq T^{\pi,\Omega}V_2$

   $$T^{\pi,\Omega}V_1 - T^{\pi,\Omega}V_2 = T^{\pi}V_1-T^{\pi}V_2\succeq\vec{0}$$

   - Distributivity: $T^{\pi,\Omega}(V+c\vec{1}) = T^{\pi,\Omega}(V)+\gamma c\vec{1}$

   $$\begin{aligned}
   T^{\pi,\Omega}(V+c\vec{1}) =& T^{\pi}(V+c\vec{1})-(1-\gamma)\Omega(\pi)\\
   =& T^{\pi}(V)+\gamma c\vec{1}-(1-\gamma)\Omega(\pi) = T^{\pi\Omega}V+\gamma c\vec{1}
   \end{aligned}$$

   - Contraction: $\|T^{\pi,\Omega}V_1 - T^{\pi,\Omega}V_2\|_\infty\le\gamma\|V_1-V_2\|_\infty$

   $$\|T^{\pi,\Omega}V_1 - T^{\pi,\Omega}V_2\|_\infty = \|T^{\pi}V_1-T^{\pi}V_2\|_\infty\le\gamma\|V_1-V_2\|_\infty$$

- $T^{\pi,\Omega}$'s unique fixed point is $V^{\pi,\Omega}$;

$$\begin{aligned}
T^{\pi,\Omega}V^{\pi,\Omega} =& T^{\pi}V^{\pi,\Omega} - (1-\gamma)\Omega(\pi) \\
=& T^{\pi}\left(V^{\pi} - \Omega(\pi)\right) - (1-\gamma)\Omega(\pi) \\
=& T^{\pi}(V^{\pi}) - \gamma\Omega(\pi) - (1-\gamma)\Omega(\pi) \\
=& V^{\pi} - \Omega(\pi) = V^{\pi,\Omega}
\end{aligned}$$

5. Regularized optimal Bellman operator: $T^{*,\Omega}V = \max_{\pi\in\Pi^{MR}} T^{\pi,\Omega}V$;

$$T^{*,\Omega}V = \max_{\pi\in\Pi^{MR}} \langle \pi_s, Q_V(s,\cdot)\rangle - (1-\lambda)\Omega(\pi_s) = \Omega^*_{\gamma}(Q_V(s,\cdot))$$

- Monotonicity: $V_1 \succeq V_2 \Rightarrow T^{*,\Omega}V_1 \succeq T^{*,\Omega}V_2$.
Let V_1's optimal policy be π_1, and V_2's be π_2.

$$\begin{aligned}
T^{*,\Omega}V_1 - T^{*,\Omega}V_2 = & \max_{\pi\in\Pi^{MR}} T^{\pi,\Omega}V_1 - \max_{\pi\in\Pi^{MR}} T^{\pi,\Omega}V_2 \\
\succeq & T^{\pi_2,\Omega}V_1 - T^{\pi_2,\Omega}V_2 \succeq P^{\pi_2}(V_1 - V_2) \succeq \vec{0}
\end{aligned}$$

- Distributivity: $T^{*,\Omega}(V + c\vec{1}) = T^{*,\Omega}V + \gamma c\vec{1}$.
- Contraction: $\|T^{*,\Omega}V_1 - T^{*,\Omega}V_2\|_{\infty} \preceq \gamma\|V_1 - V_2\|_{\infty}$

$$\|T^{*,\Omega}V_1 - T^{*,\Omega}V_2\|_{\infty} \leq \|T^{\pi_1,\Omega}V_1 - T^{\pi_1,\Omega}V_2\|_{\infty} \leq \|T^{\pi_1}V_1 - T^{\pi_1}V_2\|_{\infty} \leq \gamma\|V_1 - V_2\|_{\infty}$$

- $T^{*,\Omega}$'s unique fixed point is $V^{*,\Omega}$. (We talk about sup instead of min)
First we proof $V \succeq T^{*,\Omega}V \Rightarrow V \succeq V^{*,\Omega}$:

$$\begin{aligned}
&\forall\pi, \quad V \succeq \sup_{\pi'\in\Pi^{MR}} T^{\pi',\Omega}V \succeq r^{\pi} + \gamma P^{\pi}V - (1-\gamma)\Omega(\pi) \\
\Rightarrow & V \succeq (I - \gamma P^{\pi})(r^{\pi} - (1-\gamma)\Omega(\pi)) = V^{\pi,\Omega} \qquad \Rightarrow V \succeq V^{*,\Omega}
\end{aligned}$$

Second we proof $V \preceq T^{*,\Omega}V \Rightarrow V \preceq V^{*,\Omega}$: By definition of sup,

$$\forall\epsilon, \exists\pi\in\Pi^{MR}, V \preceq T^{\pi,\Omega}V + \epsilon\cdot\vec{1} \Rightarrow V \preceq (I - \lambda P^{\pi})^{-1}[r^{\pi} - (1-\gamma)\Omega(\pi) + \epsilon\cdot\vec{1}]$$

$$V \preceq (I - \lambda P^{\pi})^{-1}[r^{\pi} - (1-\gamma)\Omega(\pi)] + \frac{\epsilon}{1-\gamma}\vec{1} \preceq V^{*,\Omega} + \frac{\epsilon}{1-\gamma}\vec{1}$$

6. Assume that $\Omega_L \leq \Omega \leq \Omega_U$, then $V^{\pi} - \Omega_U \leq V^{\pi,\Omega} \leq V^{\pi} - \Omega_L$.

$$\max_{\pi\in\Pi^{MR}} V^{\pi} - \Omega_U \leq \max_{\pi\in\Pi^{MR}} V^{\pi,\Omega} \leq \max_{\pi\in\Pi^{MR}} V^{\pi} - \Omega_L \Rightarrow V^* - \Omega_U \leq V^{*,\Omega} \leq V^* - \Omega_L$$

Furthermore,

$$\begin{aligned}
& V^* \leq V^{*,\Omega} + \Omega_U = V^{\pi^{*,\Omega},\Omega} + \Omega_U \leq V^{\pi^{*,\Omega}} + \Omega_U - \Omega_L \\
\Rightarrow & V^* - (\Omega_U - \Omega_L) \leq V^{\pi^{*,\Omega}} \leq V^*
\end{aligned}$$

2 Negative entropy

A classical example is the negative entropy $\Omega(\pi_s) = (1-\gamma)^{-1}\sum_a \pi_s(a)\ln \pi_s(a)$.

$$\Omega^*_\gamma(q_s) = \max_{\pi\in\Pi^{MR}} \langle \pi_s, q_s\rangle - \sum_a \pi_s(a)\ln \pi_s(a)$$

We change it into

$$\begin{aligned}
-\Omega^*_\gamma(q_s) &= \min_{\pi_s\succeq\vec{0}} \max_{\alpha\neq 0} \alpha\left(\sum_a \pi_s(a) - 1\right) - \langle \pi_s, q_s\rangle + \sum_a \pi_s(a)\ln \pi_s(a) \\
&= \max_{\alpha\neq 0} \min_{\pi_s\succeq\vec{0}} \alpha\left(\sum_a \pi_s(a) - 1\right) - \langle \pi_s, q_s\rangle + \sum_a \pi_s(a)\ln \pi_s(a) \\
&\Rightarrow \alpha - q_s(a) + \ln \pi_s(a) + 1 = 0, \quad \sum_a \pi_s(a) = 1 \\
&\Rightarrow \sum_a \exp\{-1 + q_s(a) - \alpha\} = 1 \Rightarrow \alpha + 1 = \ln \sum_a \exp\{q_s(a)\} \\
&\Rightarrow \pi_s(a) = \frac{\exp\{q_s(a)\}}{\sum_a \exp\{q_s(a)\}}
\end{aligned}$$

$$\Omega^*_\gamma(q_s) = \ln \sum_a \exp q_s(a) \Rightarrow \nabla\Omega^*_\gamma(q_s) = \frac{\exp\{q_s(a)\}}{\sum_a \exp\{q_s(a)\}} = \pi^*_s(a)$$

3 Regularized Modified Policy Iteration