A Theory of Regularized Markov Decision Processes

Peng Lingwei

September 18, 2019

.

Related Works

- Trust Region Policy Optimization: Policy iteration, KL penalty;
- Dynamic Policy Programming: Value iteration, KL penalty;
- Soft Q-learning: Value iteration, Shannon entropy penalty;
- Soft Actor Critic: Policy iteration, KL penalty.

They propose a general theory of regularized Markov Decision Processes that generalizes these approaches in two directions:

- Consider a larger class of regularizers;
- Consider the general modified policy iteration approach, encompassing both policy iteration and value iteration.

Background

Unregularized MDPs:

- *Model* : $\{\mathcal{S}, \mathcal{A}, \underbrace{\mathcal{R}(s,a), \mathcal{P}(s'|s,a)}_{Markovian}, \gamma\}$;
- Markov Random Policy: $\pi(\cdot|s)$;
- Criterion: $V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\{\sum_{t=0}^{\infty}\gamma^t R(s_t, a_t)|s_0 = s\}$,
  Optimal value $V^* = \max_\pi V^\pi$;
- Bellman Operation:
  $(T_\pi V)(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left\{R(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}V\right\}$;
- Q value: $Q(s,a) = R(s,a) + \gamma\mathbb{E}_{P(s'|s,a)}V(s)$;
- $T_\pi V = \langle\pi(s), Q(s,\cdot)\rangle$;
- Bellman Optimality Operation: $TV = \max_\pi T_\pi V$;
- Greedy Policy: $\pi' \in G(V) = \arg\max_\pi T_\pi V$.

Legendre-Fenchel transform: Let $\Omega : \Delta_A \to \mathbb{R}$ be a strongly convex function:

$$\forall Q_s \in \mathbb{R}^A, \Omega^*(Q_s) = \max_{\pi_s\in\Delta_A}\langle\pi_s, Q_s\rangle - \Omega(\pi_s)$$

Regularized MDPs

- *Model* : $\{\mathcal{S}, \mathcal{A}, \underbrace{\mathcal{R}(s,a), \mathcal{P}(s'|s,a)}_{Markovian}, \gamma, \Omega\}$;
- Markov Random Policy: $\pi(\cdot|s)$;
- Criterion:
  $V^{\pi,\Omega}(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\{\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)|s_0 = s\} - \Omega(\pi)$;

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

Regularized MDPs

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

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

- Optimal value:

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

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

Regularized Bellman Operation

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

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

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

- Monotonicity: $V_1 \succeq V_2 \Rightarrow T^{\pi,\Omega}V_1 \succeq T^{\pi,\Omega}V_2$
- Distributivity: $T^{\pi,\Omega}(V + c\vec{1}) = T^{\pi,\Omega}(V) + \gamma c\vec{1}$
- Contraction: $\|T^{\pi,\Omega}V_1 - T^{\pi,\Omega}V_2\|_\infty \leq \gamma\|V_1 - V_2\|_\infty$
- $T^{\pi,\Omega}$'s unique fixed point is $V^{\pi,\Omega}$;

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

Regularized Bellman Optimality Operation

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

- Monotonicity: $V_1 \succeq V_2 \Rightarrow T^{*,\Omega}V_1 \succeq T^{*,\Omega}V_2$.
- Distributivity: $T^{*,\Omega}(V + c\vec{1}) = T^{*,\Omega}V + \gamma c\vec{1}$.
- Contraction: $\| T^{*,\Omega}V_1 - T^{*,\Omega}V_2\|_\infty \preceq \gamma \| V_1 - V_2\|_\infty$
- $T^{*,\Omega}$'s unique fixed point is $V^{*,\Omega}$. (We talk about sup instead of min)

$$V^{*,\Omega} = T^{*,\Omega}V^{*,\Omega}.$$

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

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

$$-\Omega_\gamma^*(q_s) = \min_{\pi_s \succeq \vec{0}} \max_{\alpha \neq 0} \alpha \left(\sum_a \pi_s(a) - 1 \right) - \langle \pi_s, q_s \rangle + \sum_a \pi_s(a) \ln \pi_s(a)$$

$$\Rightarrow \pi_s(a) = \frac{\exp\{q_s(a)\}}{\sum_a \exp\{q_s(a)\}}$$

$$\Omega_\gamma^*(q_s) = \ln \sum_a \exp q_s(a) \Rightarrow \nabla \Omega_\gamma^*(q_s) = \frac{\exp\{q_s(a)\}}{\sum_a \exp\{q_s(a)\}} = \pi_s^*(a)$$

From Dynamic Programming to RMPI

1. Value Iteration:
   $\pi_{t+1} = \arg\max_\pi T_\pi V_t, V_{t+1} = T_{\pi_{t+1}} V_t; \quad (V_{t+1} = TV_t)$
2. Policy Iteration:
   $\pi_{t+1} = \arg\max_\pi T_\pi V_t, V_{t+1} = V^{\pi_t} = T^\infty_{\pi_{t+1}} V_t;$
3. Modified Policy Iteration:
   $\pi_{t+1} = \arg\max_\pi T_\pi V_t, V_{t+1} = T^m_{\pi_{t+1}} V_t.$

Regularized Modified Policy Iteration:

$$\begin{cases} \pi_{k+1} = \arg\max_\pi T_{\pi,\Omega} V_t, \\ V_{k+1} = T^m_{\pi_{k+1},\Omega} V_k \end{cases}$$