Distributionally Robust Reinforcement Learning

Peng Lingwei

September 22, 2019

Contents

1 Distributionally robust policy evaluation

Definition 1. ***(Adversarial Bellman operator).***

$$\mathcal{U}_\epsilon(\pi) = \left\{\tilde{\pi} \in \Pi^{MR} | D_{KL}(\tilde{\pi}(\cdot|s) \| \pi(\cdot|s)) \leq \epsilon(s), \forall s \in S\right\} \Rightarrow T_{\pi^\epsilon} V := \min_{\tilde{\pi} \in \mathcal{U}_\epsilon(\pi)} T^{\tilde{\pi}} V$$

Target:

$$T_{\pi^\epsilon} V = \min_{\tilde{\pi} \in \Pi^{MR}} \max_{\lambda_s > 0} [T_{\tilde{\pi}}(s)] + \lambda_s D_{KL}(\tilde{\pi}(s) \| \pi(s)) - \lambda_s \epsilon(s)$$

$$T = \max_{\tilde{\pi} \in \Pi^{MR}} \min_{\alpha \neq 0} \alpha(\sum_a \tilde{\pi}(s,a) - 1) - \langle \tilde{\pi}(s), Q_V(s,\cdot) \rangle - \lambda_s \left(\sum_a \tilde{\pi}(s,a) \ln \tilde{\pi}(s,a) - \sum_a \tilde{\pi}(s,a) \ln(\pi(s,a)) \right)$$

$$\Rightarrow \alpha - Q_V(s,\cdot) - \lambda_s (\ln \tilde{\pi}(s,\cdot) + 1 - \ln \pi(s,\cdot)) = 0 \quad (\Rightarrow T = \lambda_s - \alpha)$$

$$\Rightarrow 1 - \frac{\alpha}{\lambda_s} = \ln \sum_a \exp\{-Q_V(s,a)/\lambda_s\} \pi(s,a) \quad \left(\Rightarrow T = \lambda_s \ln \left\{ \sum_a \exp\{-Q_V(s,a)/\lambda_s\} \pi(s,a) \right) \right\}$$

$$\Rightarrow T^{\pi^\epsilon} V(s) = \max_{\lambda_s > 0} (-T) - \lambda_s \epsilon(s) = -\left\{ \min_{\lambda_s > 0} \lambda_s \ln \left\{ \sum_a \exp\{-Q_V(s,a)/\lambda_s\} \pi(s,a) \right\} + \lambda_s \epsilon(s) \right\}$$

$$\Rightarrow \pi^\epsilon(a|s,\lambda_s) \propto \sum_a \exp\{-Q_V(s,a)/\lambda_s\} \pi(s,a)$$

Furthermore, we obtain

$$\lambda_s^* = \arg\min_{\lambda_s > 0} \lambda_s \Omega^*_{\pi_s} (-Q_V/\lambda_s) + \lambda_s \epsilon(s)$$

Now, we have Adversarial Modified Policy Iteration algorithm:

$$\pi_{t+1} = \arg\min_\pi T_\pi V_t, \quad V_{t+1} = T^m_{\pi^{\epsilon_t}_{t+1}} V_t$$

Definition 2. ***(Distributionally robust modified policy iteration).***

$$\begin{cases} \epsilon_t(s) = C n_t(s)^{-\eta} \cdot 1_{\{n_t(s) \geq t/S\}} \\ \pi_{t+1} = \arg\min_\pi T_\pi V_t, \quad V_{t+1} = T^m_{\pi^{\epsilon_t}_{t+1}} V_t \end{cases}$$

Algorithm 1 Distributionally Robust Policy Iteration

Require: $C, \eta > 0$.
- **repeat**:
  - $\pi = \arg\min_\pi T_\pi V$
  - $\epsilon = C n^{-\eta}$
  - $\lambda = \arg\min_{\lambda > 0} \lambda \Omega^*_{\pi_s} (-Q_V/\lambda) + \lambda \epsilon(s)$
  - $\pi^\epsilon(s) \propto \exp\{-Q_V(s)/\lambda\} \pi(s)$
  - $V = T^{\pi^\epsilon} V$
- **until** convergence

2 Extension to regularized policies

Definition 3. ***(Soft adversarial Bellman operator).***

$$T_{\pi^\epsilon,\Omega}V = \min_{\tilde{\pi}\in\mathcal{U}_\epsilon(\pi)} T_{\tilde{\pi},\Omega}V$$

$$T_{\pi^\epsilon,\Omega}V(s) = \min_{\tilde{\pi}} \max_{\lambda_s>0} \langle \tilde{\pi}(s), Q_V(s)\rangle - \Omega(\tilde{\pi}(s)) + \lambda_s D_{KL}(\tilde{\pi}(s)\|\pi(s)) - \lambda_s\epsilon(s)$$

1. If we use $\Omega(\tilde{\pi}(s)) = \beta_s \sum_a \tilde{\pi}(s,a)\ln\tilde{\pi}(s,a)$, then,

$$\begin{aligned}
T = &\max_{\tilde{\pi}} \min_{\alpha\neq 0} \alpha(\sum_a \tilde{\pi}(s,a) - 1) - \langle \tilde{\pi}(s), Q_V(s)\rangle + \beta_s \sum_a \tilde{\pi}(s,a)\ln\tilde{\pi}(s,a) \\
&- \lambda_s \left\{ \sum_a \tilde{\pi}(s,a)\ln\tilde{\pi}(s,a) - \sum_a \tilde{\pi}(s,a)\ln\pi(s,a) \right\} \\
\Rightarrow &\alpha - Q_V(s) + \beta_s(1+\ln\tilde{\pi}(s)) - \lambda_s\{1+\ln\tilde{\pi}(s) - \ln\pi(s)\} = 0 \quad (T = -\alpha - \beta_s + \lambda_s) \\
\Rightarrow &\frac{\alpha}{\beta_s - \lambda_s} + 1 = \ln \sum_a \exp\left\{ \frac{Q_V(s)}{\beta_s-\lambda_s} - \frac{\lambda_s \ln\pi(s)}{\beta_s - \lambda_s} \right\} \\
(\Rightarrow &T = (\lambda_s - \beta_s)\ln\sum_a \exp\left\{ \frac{Q_V(s)}{\beta_s-\lambda_s} - \frac{\lambda_s \ln\pi(s)}{\beta_s - \lambda_s} \right\}) \\
\Rightarrow &\pi^\epsilon(a|s,\lambda) \propto \exp\left\{ \frac{Q_V(s)}{\beta_s-\lambda_s} - \frac{\lambda_s \ln\pi(s)}{\beta_s - \lambda_s} \right\} = \left\{ \exp\left\{ -\frac{Q_V(s)}{\lambda_s} \right\} \pi(s) \right\}^{\frac{\lambda_s}{\lambda_s-\beta_s}} \\
\Rightarrow &- T_{\pi^\epsilon,\Omega}V(s) = \min_{\lambda_s>0} (\lambda_s - \beta_s)\ln\sum_a \exp\left\{ \frac{Q_V(s)}{\beta_s-\lambda_s} - \frac{\lambda_s \ln\pi(s)}{\beta_s - \lambda_s} \right\} + \lambda_s\epsilon(s)
\end{aligned}$$

2. If $\Omega_{\pi(s)}(\tilde{\pi}(s)) = \beta_s D_{KL}(\tilde{\pi}(s)\|\pi(s))$:

$$\begin{aligned}
T = &\max_{\tilde{\pi}\in\Pi^{MR}} \min_{\alpha\neq 0} \alpha(\sum_a \tilde{\pi}(s,a) - 1) - \langle \tilde{\pi}(s), Q_V(s,\cdot)\rangle \\
&- (\lambda_s - \beta_s)\left(\sum_a \tilde{\pi}(s,a)\ln\tilde{\pi}(s,a) - \sum_a \tilde{\pi}(s,a)\ln(\pi(s,a)) \right) \\
\Rightarrow &T = (\lambda_s - \beta_s)\ln\sum_a \exp\left\{ \frac{Q_V(s)}{\beta_s-\lambda_s} \right\}\pi(s) \\
\Rightarrow &- T_{\tilde{\pi}^\epsilon,\Omega}V(s) = \min_{\lambda_s>0} (\lambda_s - \beta_s)\ln\sum_a \exp\left\{ \frac{Q_V(s)}{\beta_s-\lambda_s} \right\}\pi(s) + \lambda_s\epsilon(s)
\end{aligned}$$

3. $\mathcal{U} = \{\sum_a \pi(s,a)\ln\pi(s,a) \geq \epsilon\}$ and $\Omega(\pi(s)) = \beta_s \sum_a \pi(s,a)\ln\pi(s,a)$

$$T_{\pi^\epsilon,\Omega}V(s) = \min_{\pi\in\Pi^{MR}} \max_{\lambda_s>0} \langle \pi(s), Q_V(s)\rangle - \beta_s\Omega(\pi(s)) + \lambda_s \sum_a \pi(s,a)\ln\pi(s,a) - \lambda_s\epsilon(s)$$

Let $T = \max_\pi \min_{a\neq 0} \alpha(\sum_a \pi(s,a) - 1) - \langle \pi(s), Q_V(s)\rangle + (\beta_s - \lambda_s)\sum_a \pi(s,a)\ln\pi(s,a)$

$$T_{\pi^\epsilon,\Omega}V(s) = \min_{\lambda_s>0} (\lambda_s - \beta_s)\ln\sum_a \exp\left\{ \frac{Q_V(s)}{\beta_s-\lambda_s} \right\} + \lambda_s\epsilon(s)$$

4. $\mathcal{U} = \{\sum_a \pi(s,a) \ln \pi(s,a) \geq \epsilon\}$ and $\Omega(\tilde{\pi}(s)) = \beta_s D_{KL}(\tilde{\pi}(s) \| \pi(s))$

$$-T_{\tilde{\pi}^\epsilon,\Omega} V(s) = \min_{\lambda_s > 0} (\lambda_s - \beta_s) \ln \sum_a \exp\left\{ \frac{Q_V(s)}{\beta_s - \lambda_s} + \frac{\beta_s \ln \pi(s)}{\beta_s - \lambda_s} \right\} + \lambda_s \epsilon(s)$$

The preceeding situations belongs to a general cases:

- $U_\epsilon(\pi_1) = \left\{ \pi \in \Pi^{MR} | D_{KL}\left(\pi(s) | \pi_1(s)\right) \leq \epsilon \right\}$
- $\Omega_{\pi_2}(\pi) = \beta_s D_{KL}(\pi(s) \| \pi_2(s))$
- $T_{\pi^\epsilon,\Omega} V(s) = \min_{\pi \in U_\epsilon(\pi_1)} T_{\pi,\Omega_{\pi_2}} V(s)$

$$\begin{aligned} T_{\pi^\epsilon,\Omega} V(s) &= \min_{\pi \in \Pi^{MR}} \max_{\lambda_s > 0} T_{\pi,\Omega_{\pi_2}} V(s) + \lambda_s D_{KL}(\pi \| \pi_1) - \lambda_s \epsilon \\ &= \max_{\lambda_s > 0} \min_{\pi \in \Pi^{MR}} T_{\pi,\Omega_{\pi_2}} V(s) + \lambda_s D_{KL}(\pi \| \pi_1) - \lambda_s \epsilon \end{aligned}$$

Let $T(s) = \max_\pi \min_{\alpha \neq 0} \alpha(\sum_\alpha \pi(s,a) - 1) - \langle \pi(s), Q_V(s) \rangle + \beta_s D_{KL}(\pi \| \pi_2) - \lambda_s D_{KL}(\pi \| \pi_1)$

$T_2(s) = \alpha - Q_V(s) + \beta_s\left(1 + \ln \pi(s) - \ln \pi_2(s)\right) - \lambda_s\left(1 + \ln \pi(s) - \ln \pi_1(s)\right) = 0$

$(\lambda_s - \beta_s) \ln \pi(s) = \alpha - Q_V(s) - (\lambda_s - \beta_s) + \lambda_s \ln \pi_1(s) - \beta_s \ln \pi_2(s)$

$$\exp\left\{1 - \frac{\alpha}{\lambda_s - \beta_s}\right\} = \sum_a \exp\left\{ \frac{-Q_V(s)}{\lambda_s - \beta_s} + \frac{\lambda_s \ln \pi_1(s) - \beta \ln \pi_2(s)}{\lambda_s - \beta_s} \right\}$$

$$T(s) = \langle T_2(s), \pi(s) \rangle - \alpha - \beta_s + \lambda_s = (\lambda_s - \beta_s) \ln \sum_a \exp\left\{ \frac{-Q_V(s,a)}{\lambda_s - \beta_s} + \frac{\lambda_s \ln \pi_1(s,a) - \beta_s \ln \pi_2(s,a)}{\lambda_s - \beta_s} \right\}$$

$$-T_{\pi^\epsilon,\Omega} V(s) = \min_{\lambda_s > 0} (\lambda_s - \beta_s) \ln \sum_a \exp\left\{ \frac{-Q_V(s,a)}{\lambda_s - \beta_s} + \frac{\lambda_s \ln \pi_1(s,a) - \beta_s \ln \pi_2(s,a)}{\lambda_s - \beta_s} \right\} + \lambda_s \epsilon(s)$$

3 KL-regularized Bellman operator

$$\Omega_\lambda^*(Q_V(s,\cdot)) = \lambda \ln \mathbb{E}_{a \sim \mu(\cdot|s)} \exp(Q_V(s,a)/\lambda)$$

Let $F(x) = \frac{1}{x} \ln \mathbb{E}_{a \sim \mu(\cdot|s)} \exp\left(Q_V(s,a)x\right)$:

- $F(0) = \lim_{x \to 0} \frac{\mathbb{E}_{a \sim \mu(\cdot|s)} Q_V(s,a) \exp(Q_V(s,a)x)}{\mathbb{E}_{a \sim \mu(\cdot|s)} \exp(Q_V(s,a)x)} = \mathbb{E}_{a \sim \mu(\cdot,s)} Q_V(s,a)$;
- $F'(0) = \lim_{x \to 0} \frac{\mathbb{E}_{a \sim \mu(\cdot|s)} Q_V^2(s,a) \exp(Q_V(s,a)x) - \{\mathbb{E}_{a \sim \mu(\cdot|s)} Q_V(s,a) \exp(Q_V(s,a)x)\}^2}{\left[\mathbb{E}_{a \sim \mu(\cdot|s)} \exp(Q_V(s,a)x)\right]^2} = Var_{a \sim \mu}(Q_V(s,a))$
- $F(x) = \mathbb{E}_{a \sim \mu}[Q_V(s,a)] + \frac{1}{2\lambda} Var_{a \sim \mu}[Q_V(s,a)] + O(\frac{1}{\lambda^2})$