The connection between policy optimization and imitation learning

Chenggang Liu

1 Introduction

For cost function learning, we are trying to learn a cost function so that the optimal distribution of trajectories matches the demo distribution. For sampling policy learning, we are trying to learn a sample policy so that the proposal distribution matches the optimal importance sampling or demo distribution.

2 Imitation learning as variational inference

Let's define the optimility probability as $p(O = 1|\xi) = \exp(-C(\xi))$ and

$$p(O = 1) = \int p(O = 1|\xi)p(\xi)$$

For imitation learning, we want to match the demo distribution with the optimal distribution. To do so, we can minimize the Kullback-Leibler divergence of the demo distribution from the optimal distribution, $D_{KL}(p_h||p^*)$. It is equivalent to maximize the log-likelihood of demo trajectory under the posterior distribution:

$$\begin{aligned}
\theta &= \arg\max_\theta \ln p(\xi|O_\theta = 1) \\
&= \arg\max_\theta \ln \frac{p(O = 1|\xi)p(\xi)}{p(O = 1)} \\
&= \arg\max_\theta [-C_\theta(\xi) + \ln(p(\xi)) - \ln(p(O = 1; \theta))] \\
&= \arg\max_\theta [-C_\theta(\xi) - \ln(p(O = 1; \theta))] \qquad (1)
\end{aligned}$$

In the third line, $p(\xi)$ doesn't depend on θ, so it can be ignored. The challenge is the $p(O = 1; \theta)$, but we know it's lower bound is:

$$\begin{aligned}
\ln P(O = 1) \geq \mathcal{L}(\theta, \phi; \xi) &:= E_{\tau\sim q_\phi}[\ln(p(O = 1, \tau))] + H(q_\phi) \\
&= E_{\tau\sim q_\phi}[\ln(p(O = 1|\tau)p(\tau))] + H(q_\phi) \\
&= E_{\tau\sim q_\phi}[-C_\theta(\tau) + \ln(p(\tau))] + H(q_\phi) \qquad (2)
\end{aligned}$$

Maximizing this ELBO will make q to get close to $p(x|O = 1)$, the optimal trajectory distribution. This procedure is also called Maximum Entropy RL. To solve the IRL imitation learning problem, we have need to solve such RL problem. This is why the IRL imitation more challenging. But fortunately, we don't need to get the exact solution to the RL problem and we can use a simple function as q to approximate $p(x|O = 1)$.

Put everything together and the IRL imitation problem is to solve:

$$\max_\phi \min_\theta E_{\xi\sim p_h}[E_{\tau\sim q_\phi}[-C_\theta(\tau) + \ln(p(\tau))] + H(q_\phi) + C_\theta(\xi)]$$