Trust Region

Yanhua Huang

Jun 2018

TRPO (Trust Region Policy Optimization) limits the KL divergence for policy optimization, which achieves stable performance. PPO (Proximal Policy Optimization) has the same benefits of trust-region optimization by clipping the probability ratio.

We will first discuss why clip operation achieves the trust region. One perspective is the changing ratio, i.e., the change between two policies for the same state-action pair is clipped. Another perspective is from combining policy gradient and Q-learning. For policy gradient with entropy bonus, π has the form $\exp(\alpha A^\pi - H^\pi)$, where A^π and H^π are advantage and entropy of π, respectively. Contrary to the dueling architecture, considering some states where the action selection is important, we can ignore V^π and H^π when calculating the probability ratio, so that clipping in $[1-\epsilon, 1+\epsilon]$ is same as truncating the $|Q^\pi - Q^\pi_{\text{old}}|$. For these transitions, PPO does trust-region optimization for Q.

Let's then discuss why KL divergence is considered as a trust measurement and applied widely to statistic machine learning. From the maximum likelihood estimation $\max \log p(x)$, mark the latent variables that can not be observed in data space as z with prior $q(z)$. Rewrite the likelihood by marginal distribution gets

$$\log p(x) = \log \int_z p(x,z) = log \int_z q(z) \frac{p(x,z)}{q(z)}. \tag{1}$$

Extract the weight term by Jensen's inequality gets

$$\log p(x) \geq \int_z q(z)(\log p(x,z) - \log q(z)) = \log p(x) - KL(q(z)||p(z|x)) - \mathbb{E}_{q(z)}[\log q(z)]. \tag{2}$$

Note that this is also the derivation process of variational inference and EM algorithm (replace the integrel symbol by sum symbol). The EM algorithm get the local optimal by decreasing $KL(q(z)||p(z|x))$ successively. In E step, EM finds the optimal $q(z)$ by fixing the estimation. In M step, EM finds the optimal estimation by fixing $q(z)$. For example, in kmeans algorithm and mixture gaussian model, the latent variables are assignments and the final solution is find by iteration based on EM.