On the Generalization Gap in Reparameterizable Reinforcement Learning

Huan Wang, Stephan Zheng, Caiming Xiong, Richard Socher

Salesforce Research

June 12, 2019

Summary

- **Reparameterize** RL to decouple randomness of the environment from the policy.
- Bring in supervised learning theory to RL thanks to reparameterization.
- Theoretical guarantees on the generalization gap in RL.
- Generalization gap is related to:
  - Number of training episodes
  - Smoothness of the environment, policy and reward
  - Discrepancy between training and test environment
  - "Complexity" of the reward and transition function class.

Reparameterization using Gumbel-max trick

- Treat each episode, $\boldsymbol{s}^i = [s_0^i, s_1^i, \ldots, s_T^i] \sim \mathcal{D}_\pi$, as a sample.
- Reparameterization using Gumbel noise $g_t^i \sim G$:

$$s_{t+1}^i = \arg\max \left[\log \mathcal{P}(s_t^i, \pi(s_t^i)) + g_t^i\right]$$

- Empirical reward:

$$\begin{aligned}\hat{\pi} &= \arg \max_{\pi \in \Pi, \boldsymbol{s}^i \sim \mathcal{D}_\pi} \frac{1}{n} \sum_i R(\boldsymbol{s}^i) \\ &= \arg \max_{\pi \in \Pi, g^i \sim G} \frac{1}{n} \sum_i R(\boldsymbol{s}^i(\pi, g^i))\end{aligned}$$

- The distribution G is static and independent of learned model $\pi \rightarrow$ **as in supervised learning**!

Generalization Gap in Episodic RL

$$\left|\frac{1}{n}\sum_{i=1}^{n} R(\boldsymbol{s}^i) - \mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}'} R(\boldsymbol{s})\right|$$

$$\leq \underbrace{\left|\frac{1}{n}\sum_{i=1}^{n} R(\boldsymbol{s}^i) - \mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}_{\hat{\pi}}} R(\boldsymbol{s})\right|}_{\epsilon_{intrinsic}} + \underbrace{\left|\mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}_{\hat{\pi}}} R(\boldsymbol{s}) - \mathbb{E}_{\boldsymbol{s}\sim\mathcal{D}'} R(\boldsymbol{s})\right|}_{\epsilon_{external}}$$

- Intrinsic Gap: internal randomness from the same "reparameterizable MDP":

$$\epsilon_{intrinsic} \leq Rad(R_{\pi,\mathcal{T},\mathcal{I}}) + O\left(c\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

- External Gap: between "different reparameterizable MDPs":

$$\epsilon_{extrinsic} \leq L_r\zeta\sum_{t=0}^{T}\gamma^t\frac{\nu^t-1}{\nu-1} + L_r\epsilon\sum_{t=0}^{T}\gamma^t\nu^t$$

More info

http://proceedings.mlr.press/v97/wang19o/wang19o.pdf

https://arxiv.org/abs/1905.12654

- Come see our poster!

  Wed Jun 12th 6:30 – 9:00 PM @ Pacific Ballroom #43

- Contact us:

huan.wang@salesforce.com
stephan.zheng@salesforce.com