

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



- ullet Treat each episode, $m{s}^i = [s^i_0, s^i_1, \dots, s^i_T] \sim \mathcal{D}_{\pi}$, 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:

$$\hat{\pi} = \arg \max_{\pi \in \Pi, \mathbf{s}^i \sim \mathcal{D}_{\pi}} \frac{1}{n} \sum_i R(\mathbf{s}^i)$$

$$= \arg \max_{\pi \in \Pi, \mathbf{g}^i \sim G} \frac{1}{n} \sum_i R(\mathbf{s}^i(\pi, \mathbf{g}^i))$$

► The distribution G is static and independent of learned model $\pi \to \mathbf{as}$ in supervised learning!

Generalization Gap in Episodic RL



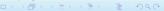
$$\left| \frac{1}{n} \sum_{i=1}^{n} R(\mathbf{s}^{i}) - \mathbb{E}_{\mathbf{s} \sim \mathcal{D}'} R(\mathbf{s}) \right| \\
\leq \underbrace{\left| \frac{1}{n} \sum_{i=1}^{n} R(\mathbf{s}^{i}) - \mathbb{E}_{\mathbf{s} \sim \mathcal{D}_{\hat{\pi}}} R(\mathbf{s}) \right|}_{\epsilon_{intrinsic}} + \underbrace{\left|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

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

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

External Gap: between "different reparameterizable MDPs":

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



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