Optimal Estimation of Policy Gradient via Double Fitted Iteration

Chengzhuo Ni[1], Ruiqi Zhang[2], Xiang Ji[1], Xuezhou Zhang[1], and Mengdi Wang[1]
[1]Department of Electrical and Computer Engineering, Princeton University
[2]School of Mathematical Science, Peking University

Markov Decision Process

- $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \xi, H)$;

$$s_1 \sim \xi, \quad s_{h+1} \sim p_h(\cdot|s_h, a_h)$$

- Policy: $\pi : \mathcal{S} \times [H] \to \Delta_{\mathcal{A}}$;
- Goal: Find the policy π to maximize $v^\pi := \mathbb{E}^\pi \left[\sum_{h=1}^H r_h(s_h, a_h)\right]$;
- Policy Gradient
  - Assume the policy has some parameterized form π_θ;

$$J(\theta) := \mathbb{E}^{\pi_\theta} \left[\sum_{h=1}^H r_h(s_h, a_h)\right].$$

  - Problem: Given a batch data $\mathcal{D}$ generated by some behavior policy $\bar{\pi}$, can we get a good estimation of $\nabla_\theta J(\theta)$?

Our Contributions

We proposed an off-policy policy gradient estimator based on double fitted iteration.

Previous works:

- The estimation error depends on the distribution shift $\max_{s\in\mathcal{S},a\in\mathcal{A},h\in[H]} \frac{\mu_h^{\pi_\theta}(s,a)}{\mu_h^{\bar{\pi}}(s,a)}$, where μ_h^{π} is the density function under policy π.

  We improved the dependency to $\mathrm{poly}(H)\max_{h\in[H]}\mathbb{E}^{\pi_\theta}\left[\frac{\mu_h^{\pi_\theta}(s,a)}{\mu_h^{\bar{\pi}}(s,a)}\right]$
- Require a known behavior policy $\bar{\pi}$.

  We don't require a known $\bar{\pi}$.
- Only apply to the tabular MDP setting.

  We consider MDP with linear features.

Approach

Suppose there is a function class $\mathcal{F}$, define

$$(\mathcal{P}_{\theta,h} f)(s,a) = \mathbb{E}^{\pi_\theta}\left[f(s_{h+1}, a_{h+1}) | s_h = s, a_h = a\right],$$

$$Q_h^\theta(s,a) = \mathbb{E}^{\pi_\theta}\left[\sum_{h'=h}^{H} r_h(s_h, a_h) | s_h = s, a_h = a\right].$$

- Policy Gradient Bellman Equation:

$$Q_h^\theta = r_h + \mathcal{P}_{\theta,h} Q_{h+1}^\theta, \ \nabla_\theta Q_h^\theta = \mathcal{P}_{\theta,h}\left(\left(\nabla_\theta \log \Pi_{\theta,h}\right) Q_{h+1}^\theta + \nabla_\theta Q_{h+1}^\theta\right),$$

  where $\left(\left(\nabla_\theta \log \Pi_{\theta,h}\right) f\right)(s,a) = \left(\nabla_\theta \log \pi_{\theta,h}(a|s)\right) f(s,a)$.

- Fitted PG Estimation:

$$\widehat{Q}_h^\theta = \arg\min_{f\in\mathcal{F}} \left[\sum_{k=1}^{K}\left(f\left(s_h^{(k)}, a_h^{(k)}\right) - r_h^{(k)} - \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a' \middle| s_{h+1}^{(k)}\right) \widehat{Q}_{h+1}^\theta\left(s_{h+1}^{(k)}, a'\right) \mathrm{d}a'\right)^2 + \lambda\rho(f)\right]$$

$$\widehat{\nabla_\theta^j Q_h^\theta} = \arg\min_{f\in\mathcal{F}} \left[\sum_{k=1}^{K}\left(f\left(s_h^{(k)}, a_h^{(k)}\right) - \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a' \middle| s_{h+1}^{(k)}\right)\left(\left(\nabla_\theta^j \log \pi_{\theta,h+1}\left(a' \middle| s_{h+1}^{(k)}\right)\right) \widehat{Q}_{h+1}^\theta\left(s_{h+1}^{(k)}, a'\right) + \widehat{\nabla_\theta^j Q_{h+1}^\theta}\left(s_{h+1}^{(k)}, a'\right)\right)\mathrm{d}a'\right)^2 + \lambda\rho(f)\right]$$

Theoretical Guarantees

- $\mathcal{F} = \{\phi(\cdot,\cdot)^\top w | w \in \mathbb{R}^d\}$ with state-action feature $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$;
- $\Sigma_h := \mathbb{E}\left[\phi\left(s_h^{(1)}, a_h^{(1)}\right)\phi\left(s_h^{(1)}, a_h^{(1)}\right)^\top\right]$, $\nu_h^\theta = \mathbb{E}^{\pi_\theta}\left[\phi(s_h, a_h)\right]$;
- $\|\nabla_\theta \pi_{\theta,h}(a|s)\|_\infty \le G, \forall s, a, h.$

Theorem (Finite Sample Guarantee)

Under certain conditions, with high probability, we have

$$\left\|\widehat{\nabla_\theta J(\theta)} - \nabla_\theta J(\theta)\right\|_\infty \le \tilde{\mathcal{O}}\left(\sqrt{\frac{H^5 G^2}{K}} \max_{h\in[H]} \left\|\Sigma_h^{-\frac{1}{2}} \nu_h^\theta\right\|\right).$$

For the tabular case,

$$\left\|\widehat{\nabla_\theta J(\theta)} - \nabla_\theta J(\theta)\right\|_\infty \le \tilde{\mathcal{O}}\left(\sqrt{\frac{H^5 G^2}{K}} \max_{h\in[H]} \mathbb{E}^{\pi_\theta}\left[\frac{\mu_h^{\pi_\theta}(s,a)}{\mu_h^{\bar{\pi}}(s,a)}\right]\right).$$

Experiments

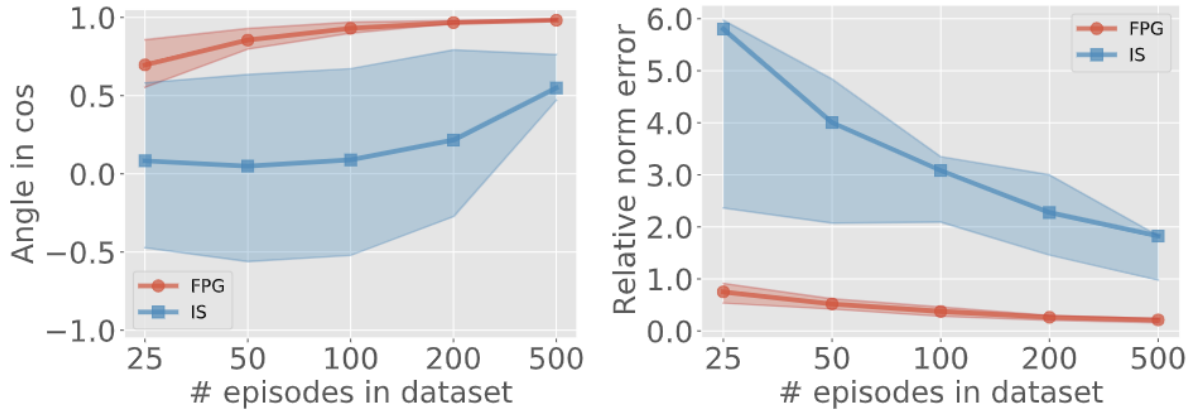

Figure 1: **Sample efficiency of FPG on off-policy data.** The off-policy PG estimation accuracy is evaluated using two metrics: $\cos\angle(\widehat{\nabla_\theta v_\theta}, \nabla_\theta v_\theta)$ and the relative error norm $\frac{\|\widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta\|}{\|\nabla_\theta v_\theta\|}$.

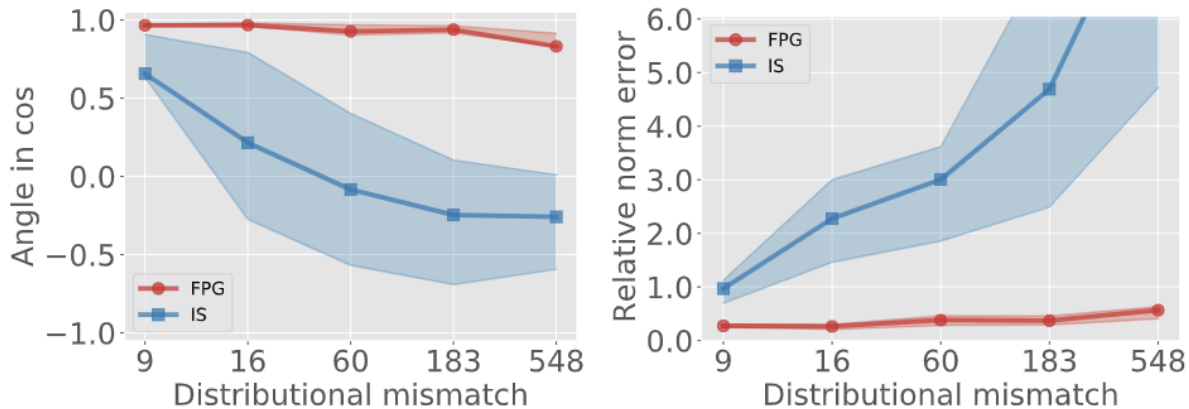

Figure 2: **Tolerance to off-policy distribution shift.** The distributional mismatch is measured by $\mathrm{cond}(\bar{\Sigma}^{\frac{1}{2}}\Sigma^{-1}\bar{\Sigma}^{\frac{1}{2}})$, where Σ is the data covariance and $\bar{\Sigma}$ is the target policy's occupancy measure.