Understanding Policy Gradient Algorithms
A Sensitivity-Based Approach

Shuang Wu[1]

Joint work w/ Ling Shi[2], Jun Wang[3], Guangjian Tian[1]

[1]**Huawei Noah's Ark Lab**, [2]Hong Kong University of Science and Technology, [3]University College London

Policy Optimization for Markov Decision Process

- Markov decision process (S, A, P, R)
  - state space S, action space A
  - state transition $P : S \times S \times A \to [0, 1]$, $P(s'|s, a)$
  - one-stage reward $R : S \times A \to \mathbb{R}$, $R(s, a)$
- Policy, stochastic: $\pi : S \times A \to [0, 1]$, deterministic: $\mu : S \to A$
- Optimization problem $\quad \max_\theta J_\bullet(\pi_\theta; \rho_0)$, $s_0 \sim \rho_0(\cdot)$
  - discounted reward $\quad J_\gamma(\pi_\theta; \rho_0) := \mathbb{E}_{\pi_\theta,\rho_0}[\sum_{k=0}^{\infty} \gamma^k R(s_k, a_k)]$
  - total reward $\quad J_{\text{tot}}(\pi_\theta; \rho_0) := \mathbb{E}_{\pi_\theta,\rho_0}[\sum_{k=0}^{\infty} R(s_k, a_k)]$
  - average reward $\quad J_{\text{av}}(\pi_\theta; \rho_0) := \lim_{T\to\infty} \mathbb{E}_{\pi_\theta,\rho_0}[\frac{1}{T+1} \sum_{k=0}^{T} R(s_k, a_k)]$

Policy Gradient $\nabla_\theta J_\bullet(\pi_\theta; \rho_0)$

In theory, [Sutton et al. 99]

$$\mathop{\mathbb{E}}_{\substack{s\sim d_\bullet^{\pi_\theta,\rho_0}(\cdot) \\ a\sim\pi_\theta(\cdot|s)}}[\nabla_\theta \log \pi_\theta(a|s) Q_\bullet^{\pi_\theta}(s,a)]$$

In practice, [Williams 88, 92]

$$\sum_{k=0}^{T} \nabla_\theta \log \pi_\theta(a_k|s_k) Q_{\text{tot}}^{\pi}(s_k, a_k)$$

Question 1: How are they related

Popular implementations, e.g., A2C/A3, ACER, ACKTR, DDPG, PPO, TD3, TRPO, and SAC, are not estimating any gradient [Nota and Thomas 20]

$$\sum_{k=0}^{T} \nabla_\theta \log \pi_\theta(a_k|s_k) Q_{\gamma}^{\pi}(s_k, a_k)$$

Question 2: How to correctly implement policy gradient without making errors?

Our Approach: Sensitivity Analysis

$$\nabla_\theta J_\bullet(\pi_\theta) = \lim_{\delta\theta \to 0} \frac{J(\pi_{\theta+\delta\theta};\rho_0) - J(\pi_\theta;\rho_0)}{\delta\theta}$$

Key observation

$$J_\bullet(\pi') - J_\bullet(\pi) = \sum_s d_\bullet^{\pi',\rho_0}(s)\Big\{ \underbrace{\mathop{\mathbb{E}}_{a\sim\pi'(\cdot|s)}\Big[Q_\bullet^\pi(s,a)\Big] - \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}\Big[Q_\bullet^\pi(s,a)\Big]}_{\text{difference between applying } \pi' \text{ and } \pi} \Big\}$$

State occupation counts
$$\begin{cases} d_\gamma^{\pi,\rho_0}(s) := & \sum_{k=0}^\infty \gamma^k \, \mathbb{E}_{s_0\sim\rho_0(\cdot)}[P^\pi(s_k = s|s_0 = s)] \\ d_{\text{tot}}^{\pi,\rho_0}(s) := & \sum_{k=0}^\infty \mathbb{E}_{s_0\sim\rho_0(\cdot)}[P^\pi(s_k = s|s_0)] \\ d_{\text{av}}^{\pi,\rho_0}(s) := & \lim_{k\to\infty} \mathbb{E}_{s_0\sim\rho_0(\cdot)}[P^\pi(s_k = s|s_0)] \end{cases}$$

Deriving Policy Gradient

$\star$ We omitted dependency on ρ_0 for brevity

Stochastic

$$
\begin{aligned}
\nabla J_\bullet(\theta) &= \lim_{\delta\theta \to 0} \frac{J_\bullet(\theta + \delta\theta) - J_\bullet(\theta)}{\delta\theta} \\
&= \lim_{\delta\theta \to 0} \sum_s d_\bullet^{\pi'}(s) \sum_a \frac{\delta\pi(a|s)}{\delta\theta} \Big(Q_\bullet^\pi(s,a) \Big) \\
&= \sum_s d_\bullet^\pi(s) \sum_a \nabla_\theta \pi(a|s) Q_\bullet^\pi(s,a) \\
&= \sum_s d_\bullet^\pi(s) \sum_a \pi(a|s) \nabla_\theta \log \pi(a|s) Q_\bullet^\pi(s,a)
\end{aligned}
$$

Deterministic

$$
\begin{aligned}
\nabla J_\bullet(\theta) &= \lim_{\delta\theta \to 0} \frac{J_\bullet(\theta + \delta\theta) - J_\bullet(\theta)}{\delta\theta} \\
&= \lim_{\delta\theta \to 0} \sum_S d_\bullet^{\mu'}(s) \left[\frac{Q_\bullet^\mu(s, \mu'(s)) - Q_\bullet^\mu(s, \mu(s))}{\delta\theta} \right] \\
&= \sum_S d_\bullet^{\mu'}(s) \left[\nabla_\theta Q_\bullet^\mu(s,a) \Big|_{a=\mu(s)} \right] \\
&= \sum_S d_\bullet^{\mu'}(s) \left[\nabla_\theta \mu(s) \nabla_a Q_\bullet^\mu(s,a) \Big|_{a=\mu(s)} \right]
\end{aligned}
$$

Extension 1: Policy Gradient in the Temporal Domain

$$\begin{cases} \nabla J_\gamma(\pi) = \sum_s d_\gamma^\pi(s) \sum_a \pi(a|s) \nabla \log \pi(a_k|s_k) Q_\gamma^\pi(s_k, a_k) \\ \nabla J_{\text{tot}}(\pi) = \sum_s d_{\text{tot}}^\pi(s) \sum_a \pi(a|s) \nabla \log \pi(a_k|s_k) Q_{\text{tot}}^\pi(s_k, a) \\ \nabla J_{\text{av}}(\pi) = \sum_s d_{\text{av}}^\pi(s) \sum_a \pi(a|s) \nabla \log \pi(a|s) Q_{\text{av}}^\pi(s_k, a) \end{cases}$$

From spatial to temporal

$$\begin{cases} d_\gamma^{\pi,\rho_0}(s) := & \sum_{k=0}^\infty \gamma^k \, \mathbb{E}_{s_0 \sim \rho_0(\cdot)}[P^\pi(s_k = s | s_0 = s)] \\ d_{\text{tot}}^{\pi,\rho_0}(s) := & \sum_{k=0}^\infty \mathbb{E}_{s_0 \sim \rho_0(\cdot)}[P^\pi(s_k = s | s_0)] \\ d_{\text{av}}^{\pi,\rho_0}(s) := & \lim_{k \to \infty} \mathbb{E}_{s_0 \sim \rho_0(\cdot)}[P^\pi(s_k = s | s_0)] \end{cases}$$

Unbiased estimates from the trajectory $\{s_0, a_0, s_1, a_1 \dots\}$

$$\begin{cases} \nabla J_\gamma(\pi) = \mathbb{E}_{s_k, a_k \sim \pi'} \left[\sum_{k=0}^\infty \gamma^k \nabla \log \pi(a_k|s_k) Q_\gamma^\pi(s_k, a_k) \right] \\ \nabla J_{\text{tot}}(\pi) = \mathbb{E}_{s_k, a_k \sim \pi'} \left[\sum_{k=0}^\infty \nabla \log \pi(a_k|s_k) Q_{\text{tot}}^\pi(s_k, a) \right] \\ \nabla J_{\text{av}}(\pi) = \lim_{T \to \infty} \mathbb{E}_{s_k \sim \pi'} \left[\frac{1}{T+1} \sum_{k=0}^T \nabla \log \pi(a|s) Q_{\text{av}}^\pi(s_k, a) \right] \end{cases}$$

Extension 2: Incorporating Policy Entropy Regularization

- Regularized single-stage reward $\tilde{R}(s,a) := R(s,a) - \tau \log \pi(a|s)$
- Value function $V_\bullet^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q_\bullet^\pi(s,a) - \tau \log \pi(a|s)]$
- Regualrized $\tilde{Q}_\bullet^\pi(s,a) := Q_\bullet^\pi(s,a) - \tau \log \pi(a|s)$
- Deriving policy gradient

$$\begin{aligned}
\nabla J_\bullet(\theta) &= \lim_{\delta\theta\to 0} \frac{J_\bullet(\theta+\delta\theta) - J_\bullet(\theta)}{\delta\theta} \\
&= \lim_{\delta\theta\to 0} \sum_s d_\bullet^{\pi'}(s) \sum_a \frac{\delta\pi(a|s)}{\delta\theta} \Big(Q_\bullet^\pi(s,a) - \tau \log \pi(a|s) \Big) \\
&\quad \underbrace{- \lim_{\delta\theta\to 0} \tau \sum_s d_\bullet^{\pi'}(\mathrm{d}s) \sum_a \frac{\pi'(\mathrm{d}a|s)}{\delta\theta} \log \frac{\pi'(a|s)}{\pi(a|s)}}_{=0} \\
&= \sum_s d_\bullet^\pi(s) \sum_a \nabla_\theta \pi(a|s) \tilde{Q}_\bullet^\pi(s,a) \\
&= \sum_s d_\bullet^\pi(s) \sum_a \pi(a|s) \nabla_\theta \log \pi(a|s) \tilde{Q}_\bullet^\pi(s,a)
\end{aligned}$$

Additional Discussion

Why current implementation still works?

- Small approximation error when $\gamma \to 1$

$$O\left(\frac{1-\gamma}{1-\alpha\gamma}\right) \text{ for some contraction factor } \alpha$$

- Maximizer Invariance
  - Experience replay changes $\max_{\pi} J(\pi;\rho_0)$ to $\max_{\pi} J(\pi;\rho_0')$
  - If the maximizer is always attainable, $\arg\max_{\pi} J(\pi;\rho_0') = \arg\max_{\pi} J(\pi;\rho_0)$

Summary

- Sensitivity analysis as a general recipe for deriving policy gradients
- Applicable for different setups (objective, regularized or not, stochastic/deterministic)
- Formal derivation of the unbiased temporal policy gradient
- Small approximation error for $\gamma \to 1$ and maximizer invariance explains empirical success