Hessian Aided Policy Gradient

Z. Shen[1], A. Ribeiro[2], H. Hassani[2], H. Qian[1], C. Mi[1]

[1]Department of Computer Science and Technology
Zhejiang University

[2]Department of Electrical and Systems Engineering
University of Pennsylvania

International Conference on Machine Learning, 2019

Outline

1. Motivation
   - Reinforcement Learning via Policy Optimization
   - Variance Reduction for Oblivious Optimization

2. Our Results/Contribution
   - Variance Reduction for Non-oblivious Optimization
   - Unbiased Policy Hessian Estimator

Outline

1. Motivation
   - Reinforcement Learning via Policy Optimization
   - Variance Reduction for Oblivious Optimization
2. Our Results/Contribution
   - Variance Reduction for Non-oblivious Optimization
   - Unbiased Policy Hessian Estimator

Policy Optimization as Stochastic Maximization

$$\max_{\theta \in \mathbb{R}^d} \mathcal{J}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{\tau \sim \pi_\theta}[\mathcal{R}(\tau)]$$

- MDP $\stackrel{\text{def}}{=} (S, A, P, r, \rho_0, \gamma)$
  $P : S \times A \times S \to [0, 1]$, $r : S \times A \to R$;
- Policy: $\pi_\theta(\cdot|s) : A \to [0, 1], \forall s \in S$;
- Trajectory:$\tau \stackrel{\text{def}}{=} (s_0, a_0, \ldots, a_{H-1}, s_H) \sim \pi_\theta$:
  $a_i \sim \pi_\theta(\cdot|s_i)$, $s_{i+1} \sim P(\cdot|s_i, a_i)$, $s_0 \sim \rho_0(\cdot)$
  Probability and discounted cumulative reward of a trajectory:

$$p(\tau) \stackrel{\text{def}}{=} \rho(s_0) \prod_{h=0}^{H-1} p(s_{h+1}|s_h, a_h)\pi_\theta(a_h|s_h)$$

$$\mathcal{R}(\tau) \stackrel{\text{def}}{=} \sum_{h=0}^{H-1} \gamma^h r(s_h, a_h)$$

Policy Optimization with REINFORCE

$$\max_{\theta \in \mathbb{R}^d} \mathcal{J}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{\tau \sim \pi_\theta}[\mathcal{R}(\tau)]$$

- Non-oblivious: $p(\tau)$ depends on θ
- REINFORCE (SGD)

$$\theta^{t+1} := \theta^t + \eta \mathbf{g}(\theta; \mathcal{S}_\tau)$$

finds $\|\mathcal{J}(\theta_\epsilon)\| \leq \epsilon$ (ϵ-FOSP) using $\mathcal{O}(1/\epsilon^4)$ samples of τ

$$\mathbf{g}(\theta; \mathcal{S}_\tau) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{S}_\tau|} \sum_{\tau \in \mathcal{S}_\tau} \mathcal{R}(\tau) \nabla \log \pi_\theta \theta(\tau), \quad \tau \in S_\tau \sim \pi_\theta$$

Outline

1. Motivation
   - Reinforcement Learning via Policy Optimization
   - Variance Reduction for Oblivious Optimization
2. Our Results/Contribution
   - Variance Reduction for Non-oblivious Optimization
   - Unbiased Policy Hessian Estimator

Oblivious Stochastic Optimization

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{z \sim p(z)}[\tilde{\mathcal{L}}(\theta; z)] \tag{1}$$

- Oblivious: $p(z)$ is independent of θ
- Stochastic Gradient Descent (SGD)

$$\theta^{t+1} := \theta^t - \eta \nabla \tilde{\mathcal{L}}(\theta^t; \mathcal{S}_z)$$

finds $\|\mathcal{L}(\theta_\epsilon)\| \leq \epsilon$ (ϵ-FOSP) using $\mathcal{O}(1/\epsilon^4)$ samples of z

$$\tilde{\mathcal{L}}(\theta; \mathcal{S}_z) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{S}_z|} \sum_{z \in \mathcal{S}_z} \tilde{\mathcal{L}}(\theta; z)$$

Variance Reduction
Oblivious Case

$$\min_{\theta\mathbb{R}^d} \mathcal{L}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{z\sim p(z)}[\tilde{\mathcal{L}}(\theta; z)] \tag{2}$$

- Oblivious: $p(z)$ is independent of θ
- SPIDER $\mathbf{g}^t := \mathbf{g}^{t-1} + \underbrace{\Delta^t \stackrel{\text{def}}{=} \left[\nabla\tilde{\mathcal{L}}(\theta^t; \mathcal{S}_z) - \nabla\tilde{\mathcal{L}}(\theta^{t-1}; \mathcal{S}_z)\right]}_{\mathbb{E}_{\mathcal{S}_z}[\Delta^t] = \nabla\mathcal{L}(\theta^t) - \nabla\mathcal{L}(\theta^{t-1})}$

$$\theta^{t+1} := \theta^t - \eta \cdot \mathbf{g}^t, \quad (\mathbb{E}[\mathbf{g}^t] = \nabla\mathcal{L}(\theta^t))$$

finds $\|\mathcal{L}(\theta_\epsilon)\| \le \epsilon$ using $\mathcal{O}(1/\epsilon^3)$ samples of z

$$\tilde{\mathcal{L}}(\theta; \mathcal{S}_z) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{S}_z|} \sum_{z\in\mathcal{S}_z} \tilde{\mathcal{L}}(\theta; z)$$

Variance Reduction

Non-oblivious Case?

$$\max_{\theta \in \mathbb{R}^d} \mathcal{J}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{\tau \sim \pi_\theta}[\mathcal{R}(\tau)] \tag{3}$$

- Non-oblivious: $p(\tau)$ depends on θ
- SPIDER

$$\mathbf{g}^t := \mathbf{g}^{t-1} + \underbrace{\Delta^t \stackrel{\text{def}}{=} \left[\mathbf{g}(\theta^t; \mathcal{S}_\tau) - \mathbf{g}(\theta^{t-1}; \mathcal{S}_\tau)\right]}_{\mathbb{E}_{\mathcal{S}_\tau}[\Delta^t] \neq \nabla \mathcal{J}(\theta^t) - \nabla \mathcal{J}(\theta^{t-1})}, \quad \tau \in S_\tau \sim \pi_{\theta^t}$$

$$\theta^{t+1} := \theta^t + \eta \mathbf{g}^t, \quad (\mathbb{E}[\mathbf{g}^t] \neq \nabla \mathcal{J}(\theta^t))$$

$$\mathbf{g}(\theta; \mathcal{S}_\tau) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{S}_\tau|} \sum_{\tau \in \mathcal{S}_\tau} \mathcal{R}(\tau) \nabla \log \pi_\theta \theta(\tau)$$

Outline

1. Motivation
   - Reinforcement Learning via Policy Optimization
   - Variance Reduction for Oblivious Optimization
2. Our Results/Contribution
   - Variance Reduction for Non-oblivious Optimization
   - Unbiased Policy Hessian Estimator

Variance Reduction for Non-oblivious Optimization

$$\theta^{t+1} := \theta^t + \eta \mathbf{g}^t, \quad (\mathbb{E}[\mathbf{g}^t] = \nabla \mathcal{J}(\theta^t))$$

- $\mathbf{g}^t := \mathbf{g}^{t-1} + \Delta^t, \mathbb{E}[\Delta^t] = \nabla \mathcal{J}(\theta^t) - \nabla \mathcal{J}(\theta^{t-1})$
- $\theta_a \stackrel{\text{def}}{=} a \cdot \theta^t + (1-a) \cdot \theta^{t-1}, a \in [0, 1]$

$$\begin{aligned} \nabla \mathcal{J}(\theta^t) - \nabla \mathcal{J}(\theta^{t-1}) &= \int_0^1 [\nabla^2 \mathcal{J}(\theta_a) \cdot (\theta^t - \theta^{t-1})]\mathbf{d}a \\ &= \left[\int_0^1 \nabla^2 \mathcal{J}(\theta_a)\mathbf{d}a\right] \cdot (\theta^t - \theta^{t-1}) \\ (\mathbb{E}_{\tau_a}[\tilde{\nabla}^2(\theta_a; \tau_a)] = \nabla^2 \mathcal{J}(\theta_a)) &= \mathbb{E}_{a \sim \mathrm{Uni}([0,1])}[\nabla^2 \mathcal{J}(\theta_a)] \cdot (\theta^t - \theta^{t-1}), \\ &= \mathbb{E}[\tilde{\nabla}^2(\theta_a) \cdot (\theta^t - \theta^{t-1})] \end{aligned}$$

Variance Reduction
Non-oblivious Case!

$$\max_{\theta\in\mathbb{R}^d} \mathcal{J}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{\tau\sim\pi_\theta}[\mathcal{R}(\tau)] \tag{4}$$

- HAPG $\mathbf{g}^t := \mathbf{g}^{t-1} + \tilde{\nabla}^2(\theta^t, \theta^{t-1}; \mathcal{S}_{a,\tau})[\theta^t - \theta^{t-1}]$

$$\theta^{t+1} := \theta^t + \eta \mathbf{g}^t, \quad (\mathbb{E}[\mathbf{g}^t] = \mathcal{J}(\theta^t))$$

$$\tilde{\nabla}^2(\theta^t, \theta^{t-1}; \mathcal{S}_{a,\tau}) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{S}_{a,\tau}|} \sum_{(a,\tau_a)\in\mathcal{S}_{a,\tau}} \tilde{\nabla}^2(\theta_a; \tau_a),$$

where $a \sim \mathrm{Uni}([0,1]), \tau_a \sim \pi_{\theta_a}.(\theta_a \stackrel{\text{def}}{=} a\cdot\theta^t + (1-a)\cdot\theta^{t-1})$

- finds $\|\mathcal{J}(\theta_\epsilon)\| \le \epsilon$ using $\mathcal{O}(1/\epsilon^3)$ samples of τ.

Outline

1. Motivation
   - Reinforcement Learning via Policy Optimization
   - Variance Reduction for Oblivious Optimization
2. Our Results/Contribution
   - Variance Reduction for Non-oblivious Optimization
   - **Unbiased Policy Hessian Estimator**

Unbiased Policy Hessian Estimator

$$\nabla \mathcal{J}(\theta) = \int_\tau \mathcal{R}(\tau)\nabla p(\tau;\pi_\theta)\mathbf{d}\tau = \int_\tau p(\tau;\pi_\theta)\cdot[\mathcal{R}(\tau)\nabla\log p(\tau;\pi_\theta)]\,\mathbf{d}\tau$$

$$\begin{aligned}
&\nabla^2\mathcal{J}(\theta)\\
&= \int_\tau \mathcal{R}(\tau)\nabla p(\tau;\pi_\theta)[\nabla\log p(\tau;\pi_\theta)]^\top + p(\tau;\pi_\theta)\cdot[\mathcal{R}(\tau)\nabla^2\log p(\tau;\pi_\theta)]\mathbf{d}\tau\\
&= \int_\tau \mathcal{R}(\tau)p(\tau;\pi_\theta)\{\nabla\log p(\tau;\pi_\theta)[\nabla\log p(\tau;\pi_\theta)]^\top + \nabla^2\log p(\tau;\pi_\theta)\}\mathbf{d}\tau
\end{aligned}$$

$$\tilde{\nabla}^2(\theta;\tau) \stackrel{\text{def}}{=} \mathcal{R}(\tau)\{\nabla\log p(\tau;\pi_\theta)[\nabla\log p(\tau;\pi_\theta)]^\top + \nabla^2\log p(\tau;\pi_\theta)\}, \tau\sim\pi_\theta.$$

Summary

First method that provably reduces the sample complexity to achieve an ϵ-FOSP of the RL objective from $\mathcal{O}(\frac{1}{\epsilon^4})$ to $\mathcal{O}(\frac{1}{\epsilon^3})$.