Hessian Aided Policy Gradient

Z. Shen\textsuperscript{1}, A. Ribeiro\textsuperscript{2}, H. Hassani\textsuperscript{2}, H. Qian\textsuperscript{1}, C. Mi\textsuperscript{1}

\textsuperscript{1}Department of Computer Science and Technology
Zhejiang University

\textsuperscript{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
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

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

- MDP \( \overset{\text{def}}{=} (S, A, P, r, \rho_0, \gamma) \)
  \[
P : S \times A \times S \to [0, 1], \quad r : S \times A \to \mathbb{R};
\]

- Policy: \( \pi_\theta(\cdot|s) : A \to [0, 1], \forall s \in S; \)

- Trajectory: \( \tau \overset{\text{def}}{=} (s_0, a_0, \ldots, a_{H-1}, s_H) \sim \pi_\theta: \)
  \[
a_i \sim \pi_\theta(\cdot|s_i), \quad s_{i+1} \sim P(\cdot|s_i, a_i), \quad s_0 \sim \rho_0(\cdot)
\]

Probability and discounted cumulative reward of a trajectory:

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

\[
R(\tau) \overset{\text{def}}{=} \sum_{h=0}^{H-1} \gamma^h r(s_h, a_h)
\]

Shen, Ribeiro, Hassani, Qian, Mi

Hessian Aided Policy Gradient
Policy Optimization with REINFORCE

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

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

\[
\theta^{t+1} := \theta^t + \eta g(\theta; S_\tau)
\]

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

\[
g(\theta; S_\tau) \overset{\text{def}}{=} \frac{1}{|S_\tau|} \sum_{\tau \in S_\tau} \mathcal{R}(\tau) \nabla \log \pi_{\theta}(\tau), \quad \tau \in S_\tau \sim \pi_{\theta}
\]
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

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

- **Oblivious**: \( p(z) \) is independent of \( \theta \)
- **Stochastic Gradient Descent (SGD)**

\[
\theta^{t+1} := \theta^t - \eta \nabla \tilde{\mathcal{L}}(\theta^t; S_z)
\]

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

\[
\tilde{\mathcal{L}}(\theta; S_z) \overset{\text{def}}{=} \frac{1}{|S_z|} \sum_{z \in S_z} \tilde{\mathcal{L}}(\theta; z)
\]
Variance Reduction
Oblivious Case

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

- **Oblivious:** \( p(z) \) is independent of \( \theta \)
- **SPIDER** \( g^t := g^{t-1} + \Delta^t \overset{\text{def}}{=} \left[ \nabla \tilde{\mathcal{L}}(\theta^t; S_z) - \nabla \tilde{\mathcal{L}}(\theta^{t-1}; S_z) \right] \)

\[ \mathbb{E}_{S_z}[\Delta^t] = \nabla \mathcal{L}(\theta^t) - \nabla \mathcal{L}(\theta^{t-1}) \]

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

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

\[
\tilde{\mathcal{L}}(\theta; S_z) \overset{\text{def}}{=} \frac{1}{|S_z|} \sum_{z \in S_z} \tilde{\mathcal{L}}(\theta; z)
\]
Variance Reduction
Non-oblivious Case?

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

- Non-oblivious: \( p(\tau) \) depends on \( \theta \)
- SPIDER
  \[
  g^t := g^{t-1} + \Delta^t \overset{\text{def}}{=} \left[ g(\theta^t; S_{\tau}) - g(\theta^{t-1}; S_{\tau}) \right], \quad \tau \in S_{\tau} \sim \pi_{\theta^t}
  \]
  \[
  \mathbb{E}_{S_{\tau}} [\Delta^t] \neq \nabla J(\theta^t) - \nabla J(\theta^{t-1})
  \]

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

\[
g(\theta; S_{\tau}) \overset{\text{def}}{=} \frac{1}{|S_{\tau}|} \sum_{\tau \in S_{\tau}} R(\tau) \nabla \log \pi_{\theta}(\tau)
\]
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

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

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

\[
\nabla J(\theta^t) - \nabla J(\theta^{t-1}) = \int_0^1 [\nabla^2 J(\theta_a) \cdot (\theta^t - \theta^{t-1})] da
\]

\[
= \left[ \int_0^1 \nabla^2 J(\theta_a) da \right] \cdot (\theta^t - \theta^{t-1})
\]

\[
(E_{\tau_a}[\tilde{\nabla}^2(\theta_a; \tau_a)] = \nabla^2 J(\theta_a)) = E_{a \sim \text{Uni}([0,1])}[\nabla^2 J(\theta_a)] \cdot (\theta^t - \theta^{t-1}),
\]

\[
= E[\tilde{\nabla}^2(\theta_a) \cdot (\theta^t - \theta^{t-1})]
\]
Motivation

Our Results/Contribution

Summary

Variance Reduction for Non-oblivious Optimization

Unbiased Policy Hessian Estimator

Variance Reduction

Non-oblivious Case!

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

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

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

\[
\tilde{\nabla}^2(\theta^t, \theta^{t-1}; S_{a,\tau}) \overset{\text{def}}{=} \frac{1}{|S_{a,\tau}|} \sum_{(a,\tau_a) \in S_{a,\tau}} \tilde{\nabla}^2(\theta_a; \tau_a),
\]

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

- finds \( \|\mathcal{J}(\theta_\epsilon)\| \leq \epsilon \) using \( \mathcal{O}(1/\epsilon^3) \) samples of \( \tau \).
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
Unbiased Policy Hessian Estimator

\[ \nabla J(\theta) = \int_\tau R(\tau) \nabla p(\tau; \pi_\theta) \, d\tau = \int_\tau p(\tau; \pi_\theta) \cdot [R(\tau) \nabla \log p(\tau; \pi_\theta)] \, d\tau \]

\[ \nabla^2 J(\theta) \]

\[ = \int_\tau R(\tau) \nabla p(\tau; \pi_\theta)[\nabla \log p(\tau; \pi_\theta)]^\top + p(\tau; \pi_\theta) \cdot [R(\tau) \nabla^2 \log p(\tau; \pi_\theta)]d\tau \]

\[ = \int_\tau R(\tau)p(\tau; \pi_\theta)\{\nabla \log p(\tau; \pi_\theta)[\nabla \log p(\tau; \pi_\theta)]^\top + \nabla^2 \log p(\tau; \pi_\theta)\}d\tau \]

\[ \tilde{\nabla}^2(\theta; \tau) \overset{\text{def}}{=} R(\tau)\{\nabla \log p(\tau; \pi_\theta)[\nabla \log p(\tau; \pi_\theta)]^\top + \nabla^2 \log p(\tau; \pi_\theta)\}, \tau \sim \pi_\theta. \]
First method that provably reduces the sample complexity to achieve an $\epsilon$-FOSP of the RL objective from $\mathcal{O}(\frac{1}{\epsilon^4})$ to $\mathcal{O}(\frac{1}{\epsilon^3})$.