From Importance Sampling to Doubly Robust Policy Gradient

Jiawei Huang (UIUC)

Nan Jiang (UIUC)

Basic Idea

Policy Gradient Estimators

Off-Policy Evaluation Estimators

Basic Idea

$$\nabla_\theta J(\pi_\theta) = \lim_{\Delta\theta \to 0} \frac{J(\pi_{\theta+\Delta\theta}) - J(\pi_\theta)}{\Delta\theta}$$

↙ Policy Gradient Estimators

↖ Off-Policy Evaluation Estimators

$$\nabla_\theta J(\pi_\theta) = \lim_{\Delta\theta \to 0} \frac{J(\pi_{\theta+\Delta\theta}) - J(\pi_\theta)}{\Delta\theta}$$

↙ ↖

Policy Gradient Estimators

REINFORCE

$$\sum_{t=0}^{T} \nabla \log \pi_\theta^t \sum_{t'=0}^{T} \gamma^{t'} r_{t'}$$

Standard PG

$$\sum_{t=0}^{T} \nabla \log \pi_\theta^t \sum_{t'=t}^{T} \gamma^{t'} r_{t'}$$

Off-Policy Evaluation Estimators

Traj-wise IS

$$\rho_{[0:T]} \sum_{t=0}^{T} \gamma^t r_t$$

Step-wise IS

$$\sum_{t=0}^{T} \gamma^t \rho_{[0:t]} r_t$$

(Tang and Abbeel, 2010)

Basic Idea

$$\nabla_\theta J(\pi_\theta) = \lim_{\Delta\theta \to 0} \frac{J(\pi_{\theta+\Delta\theta}) - J(\pi_\theta)}{\Delta\theta}$$

↙ ↖

Policy Gradient Estimators | **Off-Policy Evaluation Estimators**

PG with State Baselines

$$\sum_{t=0}^{T} \nabla \log \pi_\theta^t \Big(\sum_{t'=t}^{T} \gamma^{t'} r_{t'} - \gamma^t b_t \Big)$$

OPE with State Baselines

$$b_0 + \sum_{t=0}^{T} \gamma^t \rho_{[0:t]} \Big(r_t + \gamma b_{t+1} - b_t \Big)$$

Basic Idea

$$\nabla_\theta J(\pi_\theta) = \lim_{\Delta\theta \to 0} \frac{J(\pi_{\theta+\Delta\theta}) - J(\pi_\theta)}{\Delta\theta}$$

↙ ↖

Policy Gradient Estimators **Off-Policy Evaluation Estimators**

Trajectory-wise CV (Cheng et al., 2019)

$$\sum_{t=0}^{T} \Big\{ \nabla \log \pi_\theta^t \Big[\sum_{t'=t}^{T} \gamma^{t'} r_{t'} + \sum_{t'=t+1}^{T} \gamma^{t'} \Big(\widetilde{V}_{t'}^{\pi_\theta} - \widetilde{Q}_{t'}^{\pi_\theta} \Big) \Big] + \gamma^t \Big(\nabla \widetilde{V}_t^{\pi_\theta} - \widetilde{Q}_t^{\pi_\theta} \nabla \log \pi_\theta^t \Big) \Big\}$$

↖

Doubly Robust OPE

$$\widetilde{V}_0^{\pi'} + \sum_{t=0}^{T} \gamma^t \rho_{[0:t]} \Big(r_t + \gamma \widetilde{V}_{t+1}^{\pi'} - \widetilde{Q}_t^{\pi'} \Big)$$

↙

DR-PG **(Ours)**

$$\sum_{t=0}^{T} \Big\{ \nabla \log \pi_\theta^t \Big[\sum_{t'=t}^{T} \gamma^{t'} r_{t'} + \sum_{t'=t+1}^{T} \gamma^{t'} \Big(\widetilde{V}_{t'}^{\pi_\theta} - \widetilde{Q}_{t'}^{\pi_\theta} \Big) \Big] + \gamma^t \Big(\nabla \widetilde{V}_t^{\pi_\theta} - \nabla_\theta \widetilde{Q}_t^{\pi_\theta} - \widetilde{Q}_t^{\pi_\theta} \nabla \log \pi_\theta^t \Big) \Big\}$$

Preliminaries

MDP Setting

- Episodic RL with discount factor γ, and maximum episode length T;
- Fixed initial state distribution;
- Trajectory is defined as $s_0, a_0, r_0, s_1, ..., s_T, a_T, r_T$.

Frequently used notations

- π_θ: Policy parameterized by θ.
- $J(\pi_\theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)]$: Expected discounted return of π_θ.

From Stepwise IS OPE to Standard PG

π_θ is the behavior policy and $\pi_{\theta+\Delta\theta}$ as the target policy. $r_t = r(s_t, a_t)$ and $\pi_\theta^t = \pi_\theta(a_t|s_t)$.

$$\widehat{J}(\pi_{\theta+\Delta\theta}) = \sum_{t=0}^{T} \gamma^t r_t \prod_{t'=0}^{t} \frac{\pi_{\theta+\Delta\theta}^{t'}}{\pi_\theta^{t'}}$$

From Stepwise IS OPE to Standard PG

π_θ is the behavior policy and $\pi_{\theta+\Delta\theta}$ as the target policy. $r_t = r(s_t, a_t)$ and $\pi_\theta^t = \pi_\theta(a_t|s_t)$.

$$\widehat{J}(\pi_{\theta+\Delta\theta}) = \sum_{t=0}^{T} \gamma^t r_t \prod_{t'=0}^{t} \frac{\pi_{\theta+\Delta\theta}^{t'}}{\pi_\theta^{t'}}$$

$$= \sum_{t=0}^{T} \gamma^t r_t \Big(1 + \sum_{t'=0}^{t} \frac{\nabla_\theta \pi_\theta^{t'}}{\pi_\theta^{t'}}\Big) \Delta\theta + o(\Delta\theta)$$

From Stepwise IS OPE to Standard PG

π_θ is the behavior policy and $\pi_{\theta+\Delta\theta}$ as the target policy. $r_t = r(s_t, a_t)$ and $\pi_\theta^t = \pi_\theta(a_t|s_t)$.

$$\begin{aligned}\widehat{J}(\pi_{\theta+\Delta\theta}) &= \sum_{t=0}^{T} \gamma^t r_t \prod_{t'=0}^{t} \frac{\pi_{\theta+\Delta\theta}^{t'}}{\pi_\theta^{t'}} \\ &= \sum_{t=0}^{T} \gamma^t r_t \Big(1 + \sum_{t'=0}^{t} \frac{\nabla_\theta \pi_\theta^{t'}}{\pi_\theta^{t'}}\Big) \Delta\theta + o(\Delta\theta) \\ &= \widehat{J}(\pi_\theta) + \Big(\sum_{t=0}^{T} \gamma^t r_t \sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta^{t'}\Big) \Delta\theta + o(\Delta\theta).\end{aligned}$$

From Stepwise IS OPE to Standard PG

π_θ is the behavior policy and $\pi_{\theta+\Delta\theta}$ as the target policy. $r_t = r(s_t, a_t)$ and $\pi_\theta^t = \pi_\theta(a_t|s_t)$.

$$\begin{aligned}
\widehat{J}(\pi_{\theta+\Delta\theta}) &= \sum_{t=0}^{T} \gamma^t r_t \prod_{t'=0}^{t} \frac{\pi_{\theta+\Delta\theta}^{t'}}{\pi_\theta^{t'}} \\
&= \sum_{t=0}^{T} \gamma^t r_t \left(1 + \sum_{t'=0}^{t} \frac{\nabla_\theta \pi_\theta^{t'}}{\pi_\theta^{t'}}\right) \Delta\theta + o(\Delta\theta) \\
&= \widehat{J}(\pi_\theta) + \left(\sum_{t=0}^{T} \gamma^t r_t \sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta^{t'}\right) \Delta\theta + o(\Delta\theta).
\end{aligned}$$

Then

$$\lim_{\Delta\theta \to 0} \frac{\widehat{J}(\pi_{\theta+\Delta\theta}) - \widehat{J}(\pi_\theta)}{\Delta\theta} = \sum_{t=0}^{T} \gamma^t r_t \sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta^{t'}$$

which is known to be the standard PG.

Doubly-Robust Policy Gradient (DR-PG)

Definition: Doubly-robust OPE estimator (**unbiased**) (Jiang and Li, 2016)

$$\widehat{J}(\pi_{\theta+\Delta\theta}) = \widetilde{V}_0^{\pi_{\theta+\Delta\theta}} + \sum_{t=0}^{T} \gamma^t \Big(\prod_{t'=0}^{t} \frac{\pi_{\theta+\Delta\theta}^{t'}}{\pi_\theta^{t'}} \Big) \Big(r_t + \gamma \widetilde{V}_{t+1}^{\pi_{\theta+\Delta\theta}} - \widetilde{Q}_t^{\pi_{\theta+\Delta\theta}} \Big).$$

where $\widetilde{V}^{\theta+\Delta\theta} = \mathbb{E}_{a\sim\pi_{\theta+\Delta\theta}}[\widetilde{Q}^{\theta+\Delta\theta}]$.

Doubly-Robust Policy Gradient (DR-PG)

Definition: Doubly-robust OPE estimator (**unbiased**) (Jiang and Li, 2016)

$$\widehat{J}(\pi_{\theta+\Delta\theta}) = \widetilde{V}_0^{\pi_{\theta+\Delta\theta}} + \sum_{t=0}^{T} \gamma^t \Big(\prod_{t'=0}^{t} \frac{\pi_{\theta+\Delta\theta}^{t'}}{\pi_\theta^{t'}} \Big) \Big(r_t + \gamma \widetilde{V}_{t+1}^{\pi_{\theta+\Delta\theta}} - \widetilde{Q}_t^{\pi_{\theta+\Delta\theta}} \Big).$$

where $\widetilde{V}^{\theta+\Delta\theta} = \mathbb{E}_{a\sim\pi_{\theta+\Delta\theta}}[\widetilde{Q}^{\theta+\Delta\theta}]$.

Theorem: Given DR-OPE estimator above, we can derive two **unbiased** estimators:

- If $\widetilde{Q}^{\pi_{\theta+\Delta\theta}} = \widetilde{Q}^{\pi_\theta}$ for arbitrary $\Delta\theta$ [**Traj-CV**, (Cheng, Yan, and Boots., 2019)]

$$\sum_{t=0}^{T} \Big\{ \nabla_\theta \log \pi_\theta^t \Big[\sum_{t_1=t}^{T} \gamma^{t_1} r_{t_1} + \sum_{t_2=t+1}^{T} \gamma^{t_2} \Big(\widetilde{V}_{t_2}^{\pi_\theta} - \widetilde{Q}_{t_2}^{\pi_\theta} \Big) \Big] + \gamma^t \Big(\nabla_\theta \widetilde{V}_t^{\pi_\theta} - \widetilde{Q}_t^{\pi_\theta} \nabla_\theta \log \pi_\theta^t \Big) \Big\}.$$

- else [**DR-PG (Ours)**]

$$\sum_{t=0}^{T} \Big\{ \nabla_\theta \log \pi_\theta^t \Big[\sum_{t_1=t}^{T} \gamma^{t_1} r_{t_1} + \sum_{t_2=t+1}^{T} \gamma^{t_2} \Big(\widetilde{V}_{t_2}^{\pi_\theta} - \widetilde{Q}_{t_2}^{\pi_\theta} \Big) \Big] + \gamma^t \Big(\nabla_\theta \widetilde{V}_t^{\pi_\theta} - \nabla_\theta \widetilde{Q}_t^{\pi_\theta} - \widetilde{Q}_t^{\pi_\theta} \nabla_\theta \log \pi_\theta^t \Big) \Big\}.$$

Definition: Doubly-robust OPE estimator (Jiang and Li, 2016)

$$\widehat{J}(\pi_{\theta+\Delta\theta}) = \widetilde{V}_0^{\pi_{\theta+\Delta\theta}} + \sum_{t=0}^{T} \gamma^t \Big(\prod_{t'=0}^{t} \frac{\pi_{\theta+\Delta\theta}^{t'}}{\pi_\theta^{t'}} \Big) \Big(r_t + \gamma \widetilde{V}_{t+1}^{\pi_{\theta+\Delta\theta}} - \widetilde{Q}_t^{\pi_{\theta+\Delta\theta}} \Big).$$

Theorem: Given DR-OPE estimator above, we can derive:

- If $\widetilde{Q}^{\pi_{\theta+\Delta\theta}} = \widetilde{Q}^{\pi_\theta}$ for arbitrary $\Delta\theta$ [**Traj-CV**, (Cheng, Yan, and Boots., 2019)]

$$\sum_{t=0}^{T} \Big\{ \nabla_\theta \log \pi_\theta^t \Big[\sum_{t_1=t}^{T} \gamma^{t_1} r_{t_1} + \sum_{t_2=t+1}^{T} \gamma^{t_2} \Big(\widetilde{V}_{t_2}^{\pi_\theta} - \widetilde{Q}_{t_2}^{\pi_\theta} \Big) \Big] + \gamma^t \Big(\nabla_\theta \widetilde{V}_t^{\pi_\theta} - \widetilde{Q}_t^{\pi_\theta} \nabla_\theta \log \pi_\theta^t \Big) \Big\}.$$

- else [**DR-PG (Ours)**]

$$\sum_{t=0}^{T} \Big\{ \nabla_\theta \log \pi_\theta^t \Big[\sum_{t_1=t}^{T} \gamma^{t_1} r_{t_1} + \sum_{t_2=t+1}^{T} \gamma^{t_2} \Big(\widetilde{V}_{t_2}^{\pi_\theta} - \widetilde{Q}_{t_2}^{\pi_\theta} \Big) \Big] + \gamma^t \Big(\nabla_\theta \widetilde{V}_t^{\pi_\theta} - \nabla_\theta \widetilde{Q}_t^{\pi_\theta} - \widetilde{Q}_t^{\pi_\theta} \nabla_\theta \log \pi_\theta^t \Big) \Big\}.$$

Remark 1: The definitions of $\nabla_\theta \widetilde{V}$ are different. In Traj-CV, $\nabla_\theta \widetilde{V} = \mathbb{E}_{\pi_\theta}[\widetilde{Q}^{\pi_\theta} \nabla_\theta \log \pi_\theta]$, while in DR-PG, $\nabla_\theta \widetilde{V} = \mathbb{E}_{\pi_\theta}[\widetilde{Q}^{\pi_\theta} \nabla_\theta \log \pi_\theta + \nabla_\theta \widetilde{Q}^{\pi_\theta}]$

Remark 2: $\nabla_\theta \widetilde{Q}^{\pi_\theta}$ is not necessary a gradient but just an approximation of $\nabla_\theta Q^{\pi_\theta}$.

DR-PG

$$\sum_{t=0}^{T}\left\{\nabla_\theta \log \pi_\theta^t\left[\sum_{t_1=t}^{T}\gamma^{t_1} r_{t_1}+\sum_{t_2=t+1}^{T}\gamma^{t_2}\left(\widetilde{V}_{t_2}^{\pi_\theta}-\widetilde{Q}_{t_2}^{\pi_\theta}\right)\right]+\gamma^t\left(\nabla_\theta\widetilde{V}_t^{\pi_\theta}-\nabla_\theta\widetilde{Q}_t^{\pi_\theta}-\widetilde{Q}_t^{\pi_\theta}\nabla_\theta \log \pi_\theta^t\right)\right\}.$$

Special Cases of DR-PG

DR-PG

$$\sum_{t=0}^{T}\Big\{\nabla_\theta \log \pi_\theta^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1}r_{t_1}+\sum_{t_2=t+1}^{T}\gamma^{t_2}\Big(\widetilde{V}_{t_2}^{\pi_\theta}-\widetilde{Q}_{t_2}^{\pi_\theta}\Big)\Big]+\gamma^t\Big(\nabla_\theta\widetilde{V}_t^{\pi_\theta}-\nabla_\theta\widetilde{Q}_t^{\pi_\theta}-\widetilde{Q}_t^{\pi_\theta}\nabla_\theta\log\pi_\theta^t\Big)\Big\}.$$

Use $\widetilde{Q}^{\pi'}$ invariant to π' $\downarrow$ **Traj-CV**

$$\sum_{t=0}^{T}\Big\{\nabla_\theta \log \pi_\theta^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1}r_{t_1}+\sum_{t_2=t+1}^{T}\gamma^{t_2}\Big(\widetilde{V}_{t_2}^{\pi_\theta}-\widetilde{Q}_{t_2}^{\pi_\theta}\Big)\Big]+\gamma^t\Big(\nabla_\theta\widetilde{V}_t^{\pi_\theta}-\widetilde{Q}_t^{\pi_\theta}\nabla_\theta\log\pi_\theta^t\Big)\Big\}.$$

Special Cases of DR-PG

DR-PG

$$\sum_{t=0}^{T}\Big\{\nabla_\theta \log \pi_\theta^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1} r_{t_1} + \sum_{t_2=t+1}^{T}\gamma^{t_2}\Big(\widetilde{V}_{t_2}^{\pi_\theta} - \widetilde{Q}_{t_2}^{\pi_\theta}\Big)\Big] + \gamma^t\Big(\nabla_\theta \widetilde{V}_t^{\pi_\theta} - \nabla_\theta \widetilde{Q}_t^{\pi_\theta} - \widetilde{Q}_t^{\pi_\theta}\nabla_\theta \log \pi_\theta^t\Big)\Big\}.$$

Use $\widetilde{Q}^{\pi'}$ invariant to π' $\downarrow$ **Traj-CV**

$$\sum_{t=0}^{T}\Big\{\nabla_\theta \log \pi_\theta^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1} r_{t_1} + \sum_{t_2=t+1}^{T}\gamma^{t_2}\Big(\widetilde{V}_{t_2}^{\pi_\theta} - \widetilde{Q}_{t_2}^{\pi_\theta}\Big)\Big] + \gamma^t\Big(\nabla_\theta \widetilde{V}_t^{\pi_\theta} - \widetilde{Q}_t^{\pi_\theta}\nabla_\theta \log \pi_\theta^t\Big)\Big\}.$$

$\mathbb{E}[\sum_{t_2=t+1}^{T}\gamma^{t_2}(\widetilde{V}_{t_2}^{\pi_\theta} - \widetilde{Q}_{t_2}^{\pi_\theta})\big|s_{t+1}] = 0$, dropped $\downarrow$ **PG with state-action baselines**

$$\sum_{t=0}^{T}\Big\{\nabla_\theta \log \pi_\theta^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1} r_{t_1}\Big] + \gamma^t\Big(\nabla_\theta \widetilde{V}_t^{\pi_\theta} - \widetilde{Q}_t^{\pi_\theta}\nabla_\theta \log \pi_\theta^t\Big)\Big\}.$$

Special Cases of DR-PG

DR-PG

$$\sum_{t=0}^{T}\Big\{\nabla_\theta \log \pi_\theta^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1}r_{t_1}+\sum_{t_2=t+1}^{T}\gamma^{t_2}\Big(\widetilde{V}_{t_2}^{\pi_\theta}-\widetilde{Q}_{t_2}^{\pi_\theta}\Big)\Big]+\gamma^t\Big(\nabla_\theta\widetilde{V}_t^{\pi_\theta}-\nabla_\theta\widetilde{Q}_t^{\pi_\theta}-\widetilde{Q}_t^{\pi_\theta}\nabla_\theta \log \pi_\theta^t\Big)\Big\}.$$

Use $\widetilde{Q}^{\pi'}$ invariant to π' ↓ Traj-CV

$$\sum_{t=0}^{T}\Big\{\nabla_\theta \log \pi_\theta^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1}r_{t_1}+\sum_{t_2=t+1}^{T}\gamma^{t_2}\Big(\widetilde{V}_{t_2}^{\pi_\theta}-\widetilde{Q}_{t_2}^{\pi_\theta}\Big)\Big]+\gamma^t\Big(\nabla_\theta\widetilde{V}_t^{\pi_\theta}-\widetilde{Q}_t^{\pi_\theta}\nabla_\theta \log \pi_\theta^t\Big)\Big\}.$$

$\mathbb{E}[\sum_{t_2=t+1}^{T}\gamma^{t_2}(\widetilde{V}_{t_2}^{\pi_\theta}-\widetilde{Q}_{t_2}^{\pi_\theta})\Big|s_{t+1}]=0$, dropped ↓ **PG with state-action baselines**

$$\sum_{t=0}^{T}\Big\{\nabla_\theta \log \pi_\theta^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1}r_{t_1}\Big]+\gamma^t\Big(\nabla_\theta\widetilde{V}_t^{\pi_\theta}-\widetilde{Q}_t^{\pi_\theta}\nabla_\theta \log \pi_\theta^t\Big)\Big\}.$$

Use $\widetilde{V}$ as $\widetilde{Q}$ ↓ **PG with state baselines**

$$\sum_{t=0}^{T}\Big\{\nabla_\theta \log \pi_\theta^t\Big[\sum_{t_1=t}^{T}\gamma^{t_1}r_{t_1}\Big]+\gamma^t\Big(-\widetilde{V}_t^{\pi_\theta}\nabla_\theta \log \pi_\theta^t\Big)\Big\}.$$

Theorem The covariance matrix of the DR-PG estimator is

$$\mathbb{E}\Bigg[\sum_{n=0}^{T}\gamma^{2n}\Bigg(\underbrace{\mathbb{V}_{n+1}[r_n]\Big(\sum_{t=0}^{n}\nabla_\theta \log \pi_\theta^t\Big)\Big(\sum_{t=0}^{n}\nabla_\theta \log \pi_\theta^t\Big)^\top}_{\textit{Randomness of reward}}$$

$$+\underbrace{\mathrm{Cov}_n\Big[\nabla_\theta V_n^{\pi_\theta}+\Big(\sum_{t=0}^{n-1}\nabla_\theta \log \pi_\theta^t\Big)V_n^{\pi_\theta}\Big]}_{\textit{Randomness of transition}}$$

$$+\underbrace{\mathrm{Cov}_n\Big[\nabla_\theta Q_n^{\pi_\theta}-\boldsymbol{\nabla_\theta}\widetilde{Q}_n^{\pi_\theta}+\Big(\sum_{t=0}^{n}\nabla_\theta \log \pi_\theta^t\Big)\Big(Q_n^{\pi_\theta}-\widetilde{\mathbf{Q}}_n^{\pi_\theta}\Big)\Big|s_n\Big]}_{\textit{Randomness of policy}}\Bigg)\Bigg].$$

where

$$\begin{aligned}\mathbb{V}_n[\cdot] &:= \mathbb{V}[\cdot|s_0,a_0,\ldots s_{n-1},a_{n-1}]\\ \mathbb{E}_n[\cdot] &:= \mathbb{E}[\cdot|s_0,a_0,\ldots s_{n-1},a_{n-1}]\\ \mathrm{Cov}_n[\mathbf{v}] &:= \mathbb{E}_n[\mathbf{v}\mathbf{v}^\top]-\mathbb{E}_n[\mathbf{v}]\mathbb{E}_n[\mathbf{v}]^\top.\end{aligned}$$

Cramer-Rao Lower Bound

Theorem: For tree-structured MDPs (i.e., each state only appears at a unique time step and can be reached by a unique trajectory), the Cramer-Rao lower bound of PG is

$$\mathbb{E}\Big[\sum_{t=0}^{T}\gamma^{2t}\underbrace{\Big\{\mathbb{V}_{t+1}[r_t]\Big[\Big(\sum_{t_1=0}^{t}\frac{\partial \log \pi_\theta^{t_1}}{\partial \theta_i}\Big)\Big]^2}_{\text{Randomness of reward}}+\underbrace{\mathbb{V}_t\Big[\Big(V_t^{\pi_\theta}\sum_{t_1=0}^{t-1}\frac{\partial \log \pi_\theta^{t_1}}{\partial \theta_i}+\frac{\partial V_t^{\pi_\theta}}{\partial \theta_i}\Big)\Big]\Big\}}_{\text{Randomness of Transition}}\Big],$$

which coincides with the variance of DR-PG when $\widetilde{Q}^{\pi_\theta}\equiv Q^{\pi_\theta}$ and $\nabla_\theta\widetilde{Q}^{\pi_\theta}\equiv\nabla_\theta Q^{\pi_\theta}$.

Covariance Comparison in Special Case

Deterministic environment with perfect value function estimation

Estimator	Covariance Matrices
PG with state baselines	$\mathbb{E}\left[\sum_n \mathrm{Cov}_n\left[\nabla_\theta Q_n^{\pi_\theta} + \left(\sum_{t=0}^{n-1}\nabla_\theta \log \pi_\theta^t\right) Q_n^{\pi_\theta} + \nabla_\theta \log \pi_\theta^n A_n^{\pi_\theta} \middle\| s_n\right]\right]$
PG with state-action baselines	$\mathbb{E}\left[\sum_n \mathrm{Cov}_n\left[\nabla_\theta Q_n^{\pi_\theta} + \left(\sum_{t=0}^{n-1}\nabla_\theta \log \pi_\theta^t\right) Q_n^{\pi_\theta} \middle\| s_n\right]\right]$
Trajwise-CV	$\mathbb{E}\left[\sum_n \mathrm{Cov}_n\left[\nabla_\theta Q_n^{\pi_\theta} \middle\| s_n\right]\right]$
DR-PG	0

Experiments (Variance Reduction)

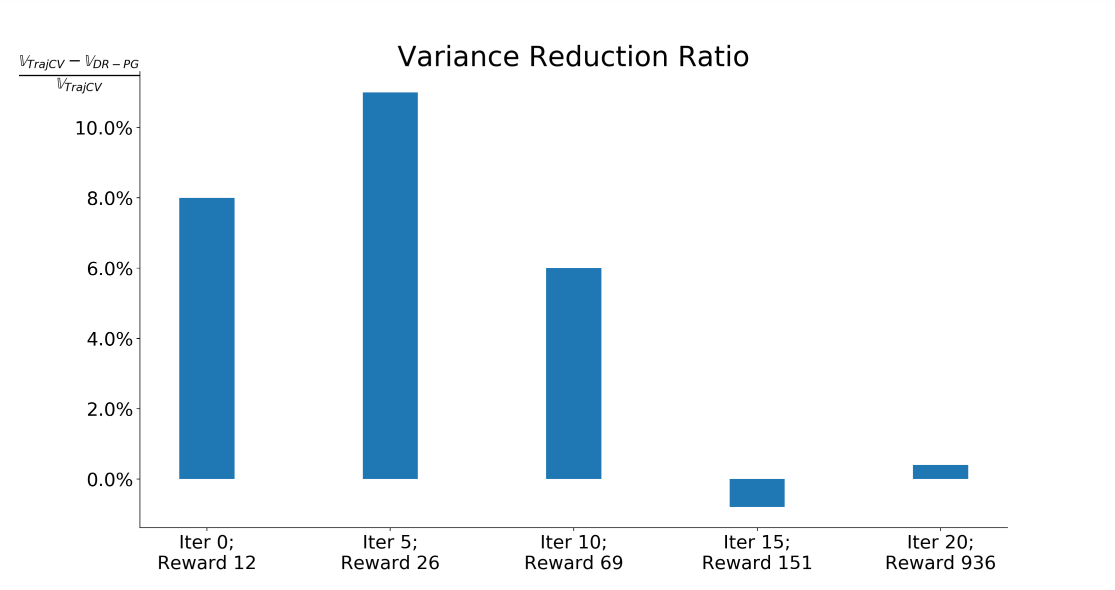

Figure 1: Variance reduction ratio. $\mathbb{V}_G$ denotes the sum of estimator G's variance over all parameters of the neural network.

Experiments (Algorithm Performance)

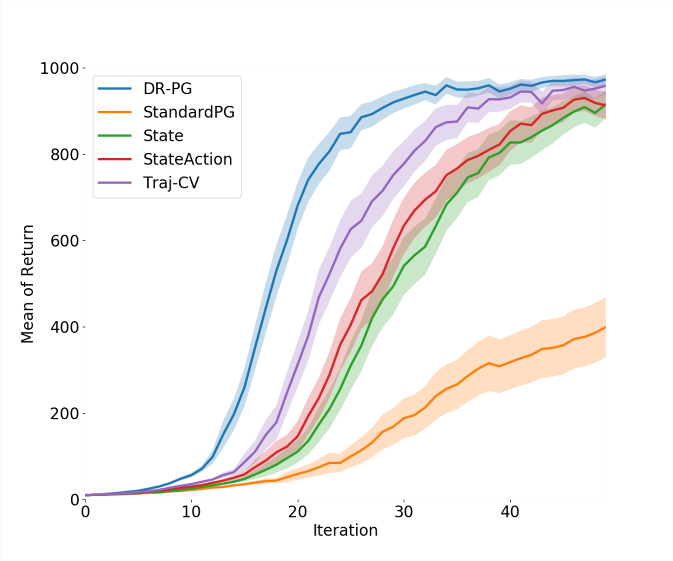

Figure 2: Performance in CartPole task. Average over 150 trials. Plot twice standard error.

Experiments (Algorithm Time Complexity)

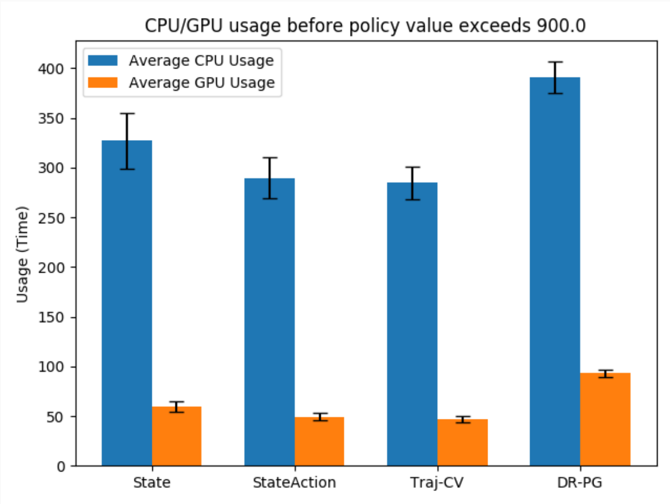

Figure 3: Comparison of GPU/CPU Usage .

Thank You!

Welcome to our Q&A session!