Variance Reduction (Part 1)

Q-prop: Sample-efficient policy gradient with an off-policy critic.

Gu, S., Lillicrap, T., Ghahramani, Z., Turner, R.E. and Levine, S., 2016. ICLT 2017

Action-dependent Control Variates for Policy Optimization via Stein's Identity

Liu, H., Feng, Y., Mao, Y., Zhou, D., Peng, J. and Liu, Q., 2017. ICLT 2018

Presented by Elena Labzina

Model-free reinforcement methods

- On-policy methods

- *Policy gradient*

- *Monte-Carlo policy gradient*

unbiased

high variance

- Off-policy methods

- *Q-learning*

- ***Off-policy critic methods***

data efficient

bad convergence/instability

Let's combine on-policy and off-policy methods' benefits!

Q-Prop (Gu et al, 2016), Policy Gradient with Stein Control Variates (Liu et al, 2017)

On-policy: Policy gradient in the reinforcement model and its estimation

$$J(\theta) = \mathbb{E}_{s \sim \rho_\pi, a \sim \pi(a|s)} [r(s, a)]$$

Expected cumulative reward

policy gradient theorem

The gradient of the expected reward

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi [\nabla_\theta \log \pi(a|s) Q^\pi(s, a)]$$

Monte Carlo estimation

$$\hat{\nabla}_\theta J(\theta) = \frac{1}{n} \sum_{t=1}^{n} \gamma^{t-1} \nabla_\theta \log \pi(a_t|s_t) \hat{Q}^\pi(s_t, a_t)$$

The estimate of the gradient

Variance reduction: Control variate

- E(s) $= \mu$ and var(s) is known and finite…. and too large!
- E(t) = 0 and var(t) is known and finite (control variate)
- s^* = s + ct
- E(s^*) = E(s) + cE(t) = μ => unbiased as well
- var(s^*) = var(s) – 2c cov(s,t) + c^2var(t)
- $\exists c$: var(s^*) $\leq$ var(s)

- argmin_c var(s^*) = cov(s,t)/var(t) (optimal c)
- var(s^*) = var(s) - cov(s,t)^2/var(t) = var(s) - corr(s,t)^{2}var(s)
- var(s^*) = var(s) (1 – corr(s,t)^2) $\leq$ var(s)

Control variate and Monte Carlo

- *Variance reduction (off-policy) technique in policy gradient*

$\mu = \mathbb{E}_\tau[g(s,a)]$

$(s_t, a_t)_{t=1}^{n}$

τ

$\mathrm{var}_\tau(g)$

control variate

$f(s,a)$

$\mathbb{E}_\tau[f(s,a)] = 0$

$$\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n}\left(g(s_t,a_t) - f(s_t,a_t)\right)$$

$\mathrm{var}_\tau(g - f)/n$

Control variate: Identification of f(a,s)

- *GOAL: develop a more general control variate with smaller variance*
- $\phi(s)$ – *base function*
- *f(a,s)* = $\nabla_\theta \log \pi(a|s)\, \phi(s)$ *(corresponding contol variate)*
- $E_{\pi(a|s)}[\nabla_\theta \log \pi(a|s)\, \phi(s)] = 0$
- *Let's modify the Monte-Carlo policy gradient:*

$$\hat{\nabla}_\theta J(\theta) = \frac{1}{n}\sum_{t=1}^{n} \gamma^{t-1} \nabla_\theta \log \pi(a_t|s_t) \hat{Q}^\pi(s_t, a_t)$$

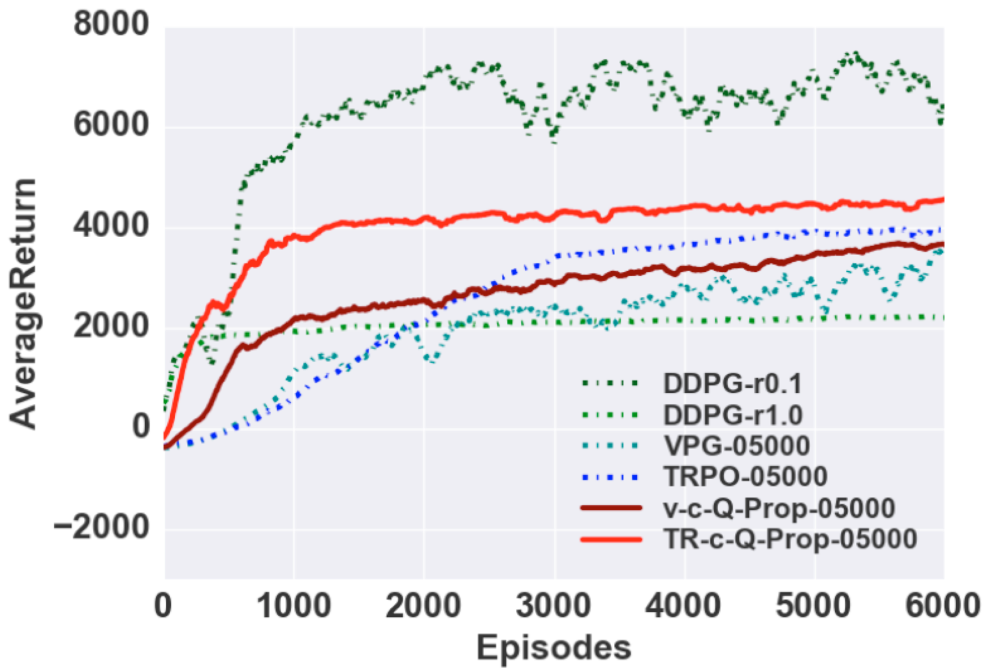
$$\hat{\nabla}_\theta J(\theta) = \frac{1}{n}\sum_{t=1}^{n} \nabla_\theta \log \pi(a_t|s_t) \left(\hat{Q}^\pi(s_t, a_t) - \phi(s_t)\right)$$

Q-Prop estimator's gist (Gu et al, 2016)

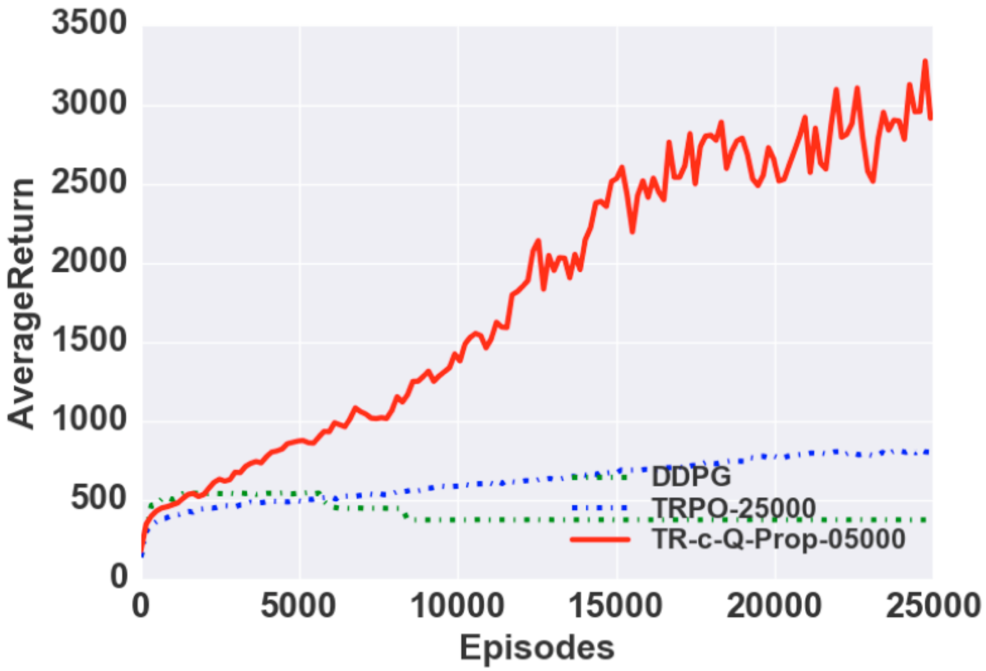
- For arbitrary function $f(s_t, a_t)$
- $\overline{f(s_t, a_t)}$ = $f(s_t, \bar{a}_t)$ + $\nabla_a f(s_t, a_t)|_{a=\bar{a}_t}(a_t - \bar{a}_t)$ (First-order Taylor expansion)
- $\overline{f(s_t, a_t)}$ = = $\phi(s)$ – base function

- $f(s_t, a_t)$ = $Q_w(s_t, a_t)$ (the critic function)
- $\bar{a}_t$ = $\mu_\Theta(s_t) = E_{\pi(a|s)}[a_t]$ (expected action of a stochastic policy π_Θ)

Q-Prop performance



(a) Comparing algorithms on HalfCheetah-v1.



(b) Comparing algorithms on Humanoid-v1.

Stein's identity for policy gradient

Given a policy $\pi(a|s)$, Stein's identity w.r.t π is

$$\mathbb{E}_{\pi(a|s)}\left[\nabla_a \log \pi(a|s)\phi(s,a) + \nabla_a \phi(s,a)\right] = 0, \qquad \forall s,$$

- $E_{\pi(a|s)}[\nabla_\theta \log \pi(a|s)\, \phi(s)] = 0$ (requirement for the control variate)
- Problem to apply!
- Given an approach to connect $\nabla_a \log \pi(a|s)$ and $\nabla_\theta \log \pi(a|s)$ any base function will work
- $a \sim \pi_\theta$(a|s) can be viewed as generated by a = f_θ(s,ξ), ξ is random noise
- $\nabla_\theta \log \pi(a,\xi|s)$ = - $\nabla_\theta\, f_\theta$(s,ξ) $\nabla_a \log \pi(a,\xi|s)$

PPO: Stein control variates vs a typical baseline

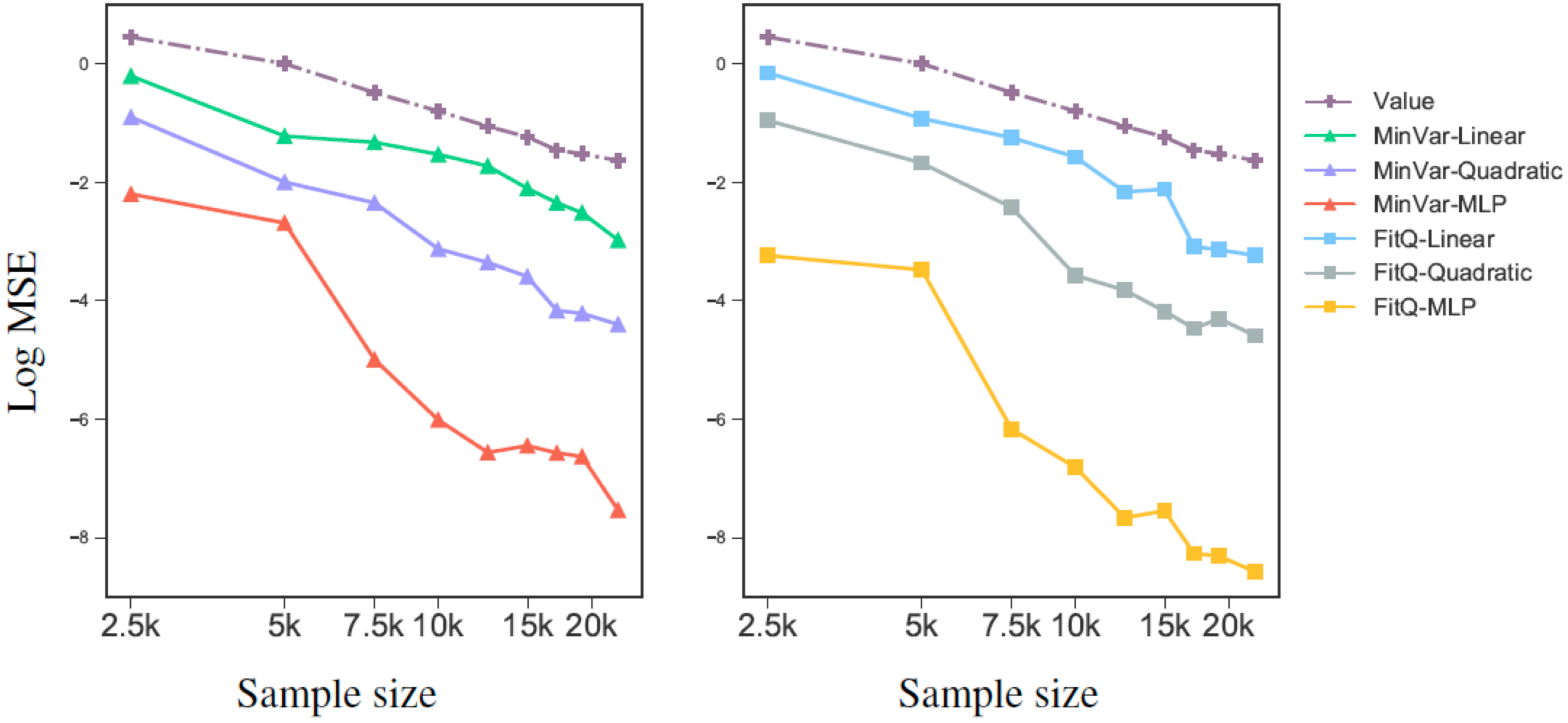
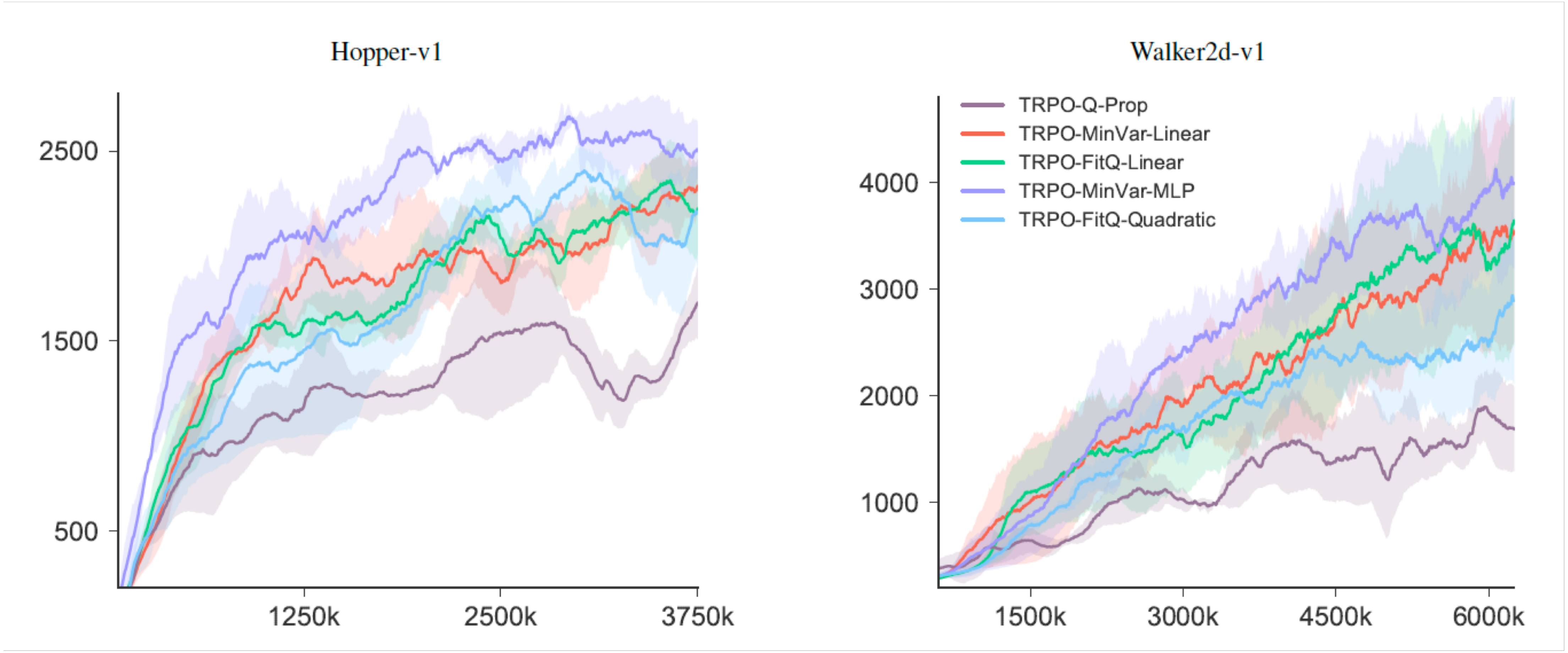
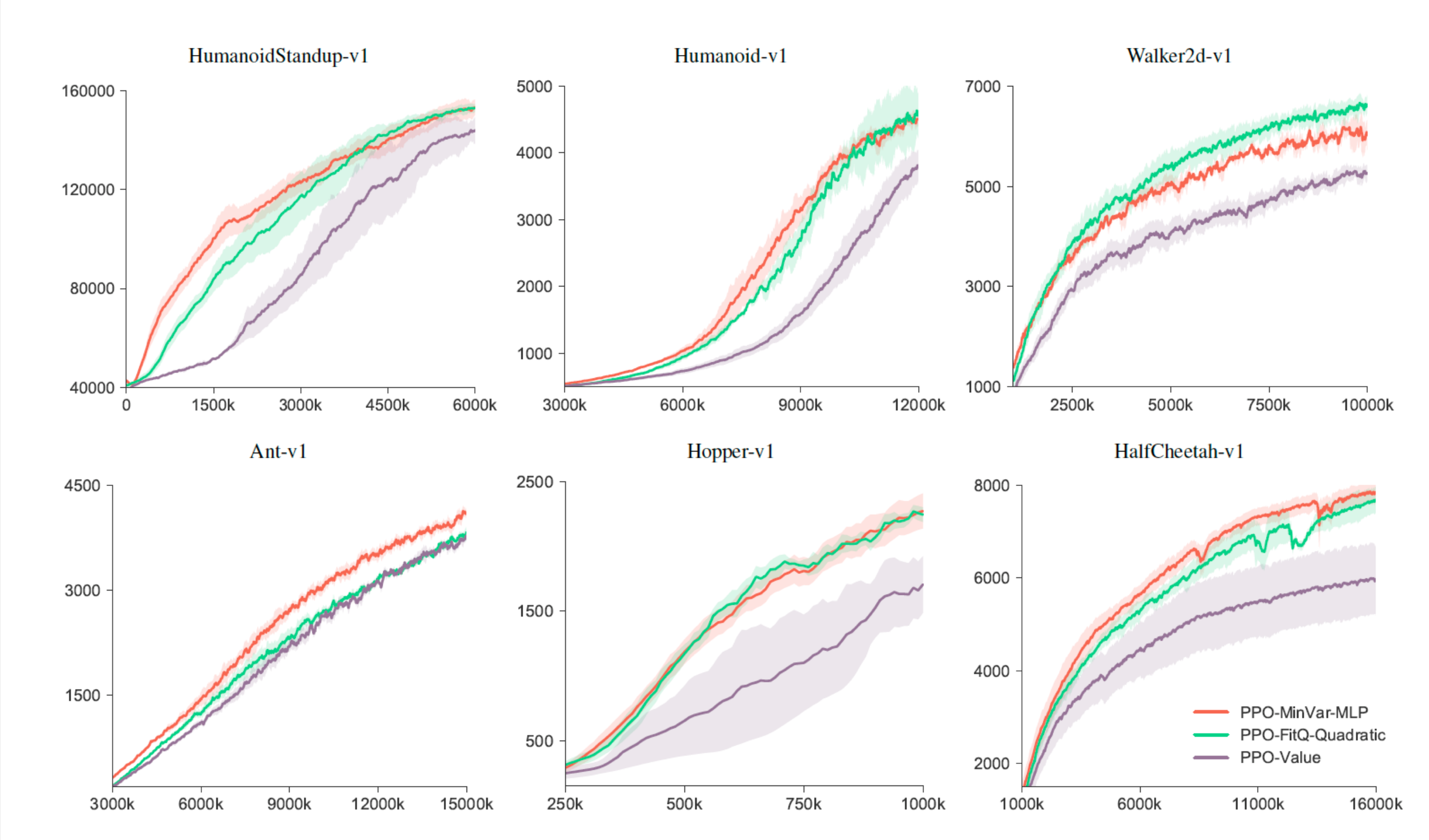


Figure 1: The variance of gradient estimators of different control variates under a fixed policy obtained by running vanilla PPO for 200 iterations in the Walker2d-v1 environment.

Evaluation of TRPO with Q-prop and Stein control variates



Evaluation of PPO with the value function baseline and Stein control variates



Take-away points

- Combination of on-policy and off-policy methods allows to need *less training data*, have *better convergence*, and have *less variance*
- *Monte Carlo policy gradient is the simplest on-policy method*
- *Control variate* is an (off-policy) method to *decrease the variance* of policy gradient methods
- *Stein Control variates* allow to create superior model-free reinforcement methods that combine on-policy and off-policy data
- *Q-prop*, *REINFORCE*, *A2C* all belong to the Stein Control variate family

THANK YOU !

Stein's identity for Monte-Carlo gradient

$$\mathbb{E}_{\pi(a|s)}\left[\nabla_a \log \pi(a|s)\phi(s,a) + \nabla_a \phi(s,a)\right] = 0, \quad \forall s,$$

Theorem

$$\mathbb{E}_{\pi(a|s)}\left[\nabla_\theta \log \pi(a|s)\phi(s,a)\right] = \mathbb{E}_{\pi(a,\xi|s)}\left[\nabla_\theta f_\theta(s,\xi)\nabla_a \phi(s,a)\right]$$

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi\left[\nabla_\theta \log \pi(a|s)(Q^\pi(s,a) - \phi(s,a)) \;+\; \nabla_\theta f_\theta(s,\xi)\nabla_a \phi(s,a)\right]$$

$$\hat{\nabla}_\theta J(\theta) = \frac{1}{n}\sum_{t=1}^{n}\Big[\nabla_\theta \log \pi(a_t \mid s_t)(\hat{Q}^\pi(s_t,a_t) - \phi(s_t,a_t)) + \nabla_\theta f_\theta(s_t,\xi_t)\nabla_a \phi(s_t,a_t)\Big]$$

Stein's identity's connection to Q-Prop

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi \left[\nabla_\theta \log \pi(a|s)(Q^\pi(s,a) - \phi(s,a)) \; + \; \nabla_\theta f_\theta(s,\xi) \nabla_a \phi(s,a)\right]$$

$$\nabla_a \phi(a, s) = \varphi(s)$$

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi \left[\nabla_\theta \log \pi(a|s)(Q^\pi(s,a) - \phi(s,a)) \; + \; \nabla_\theta f_\theta(s,\xi) \varphi(s)\right]$$

$$\mathbb{E}_{\pi(\xi)}[\nabla_\theta f(s,\xi)] = \nabla_\theta \mathbb{E}_{\pi(\xi)}[f(s,\xi)] := \nabla_\theta \mu_\pi(s).$$

$$\nabla_\theta J(\theta) = \mathbb{E}_\pi \left[\nabla_\theta \log \pi(a|s) \left(Q^\pi(s,a) - \phi(s,a)\right) \; + \; \nabla_\theta \mu_\pi(s) \varphi(s)\right]$$

$$\phi(s,a) = \hat{V}^\pi(s) + \langle \nabla_a \hat{Q}^\pi(s, \mu_\pi(s)), \; a - \mu_\pi(s) \rangle$$