Taylor Expansion of Discount Factors

Yunhao Tang, Mark Rowland, Remi Munos, Michal Valko

Motivation

- Mismatch between policy gradient theory & practice
- Theory: discounted average

$$E_\pi\left[\Sigma_{t=0}^{\infty}\gamma^t Q_\gamma^\pi(x_t,a_t)\nabla_\theta \log \pi(a_t|x_t)\right]$$

- Practical heuristic: uniform average

$$E_\pi\left[\Sigma_{t=0}^{T}1^t Q_\gamma^\pi(x_t,a_t)\nabla_\theta \log \pi(a_t|x_t)\right]$$

- Question: can we understand the gap?

Main take-away

- The discrepancy stems from the difference of objectives
- Theory studies discounted objective $V_\gamma^\pi(x)$
- Practices care about 'almost' undiscounted objective

$$E_\pi[\Sigma_{t=0}^{T} r_t | x_0 = x]$$

- Example: in MuJoCo cont control, we have $T = 1000$
- Insight: the practical heuristic can be seen as a partial gradient of the undiscounted objective

Two value functions

Examples: $\gamma = 0.99$, $T = 1000 \Rightarrow \gamma' = 0.999$

- Discounted objective with γ

$$V_{\gamma}^{\pi}(x) = E_{\pi}[\Sigma_{t=0}^{\infty} \gamma^{t} r_t | x_0 = x]$$

- Undiscounted obj over horizon T ≈ Discounted with $\gamma' = 1 - \frac{1}{T}$

$$E_{\pi}[\Sigma_{t=0}^{T} r_t | x_0 = x] \approx V_{\gamma'}^{\pi}(x) = E_{\pi}[\Sigma_{t=0}^{\infty} (\gamma')^{t} r_t | x_0 = x]$$

- What's the connection between $V_{\gamma}^{\pi}(x)$ and $V_{\gamma'}^{\pi}(x)$?

Taylor expansion of discount factors

- $V_{\gamma}^{\pi}(x)$ and $V_{\gamma'}^{\pi}(x)$ are related through Taylor expansions

Proposition 3.1. The following holds for all $K \geq 0$,

K-th order expansion in $(\gamma' - \gamma)$

$$V_{\gamma'}^{\pi} = \sum_{k=0}^{K} \left((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi}\right)^k V_{\gamma}^{\pi}$$

Residual term

$$+ \underbrace{\left((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi}\right)^{K+1} V_{\gamma'}^{\pi}}_{\text{residual}}. \quad (9)$$

When $\gamma < \gamma' < 1$, the residual norm converges to 0, which implies

Infinite series

$$V_{\gamma'}^{\pi} = \sum_{k=0}^{\infty} \left((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi}\right)^k V_{\gamma}^{\pi}. \quad (10)$$

A few properties of the expansion

- Further intuitions about the expansion: $V_{\gamma'}^{\pi}(x)$ is equivalent to

$$V_{\gamma}^{\pi}(x) + \mathbb{E}_{\pi}\left[\sum_{t=1}^{\infty}(\gamma' - \gamma)(\gamma')^{t-1}V_{\gamma}^{\pi}(x_t)\,\middle|\, x_0 = x\right]$$

'Value function' with $V_{\gamma}^{\pi}(x)$ as the reward

- K-th order approximation

$$V_{K,\gamma,\gamma'}^{\pi} := \sum_{k=0}^{K}((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1}P^{\pi})^{k}V_{\gamma}^{\pi}. \quad \Rightarrow V_{\gamma'}^{\pi}(x)$$

Can be estimated by bootstrapping with $V_{\gamma}^{\pi}(x)$

Policy gradient for $V_{\gamma'}^{\pi}$?

- Why not plug in PG formula for $V_{\gamma'}^{\pi}$?

$$E_{\pi}\left[\Sigma_{t=0}^{\infty}(\gamma')^{t}Q_{\gamma'}^{\pi}(x_t, a_t)\nabla_{\theta}\log\pi(a_t|x_t)\right]$$

- Variance might be too high, need to estimate $Q_{\gamma'}^{\pi}(x_t, a_t)$
- Need approximations

Practical heuristic as partial gradient

- The practical heuristic can be derived as a partial gradient through V_γ^π

$$E_\pi\left[\Sigma_{t=0}^{\infty}(\gamma')^t Q_\gamma^\pi(x_t, a_t)\nabla_\theta \log \pi(a_t|x_t)\right]$$

$Q_\gamma^\pi(x, a)$ can be estimated with low variance

- When $\gamma' = 1$, if the horizon is finite of length T, we derive

$$E_\pi\left[\Sigma_{t=0}^{T}1^t Q_\gamma^\pi(x_t, a_t)\nabla_\theta \log \pi(a_t|x_t)\right]$$

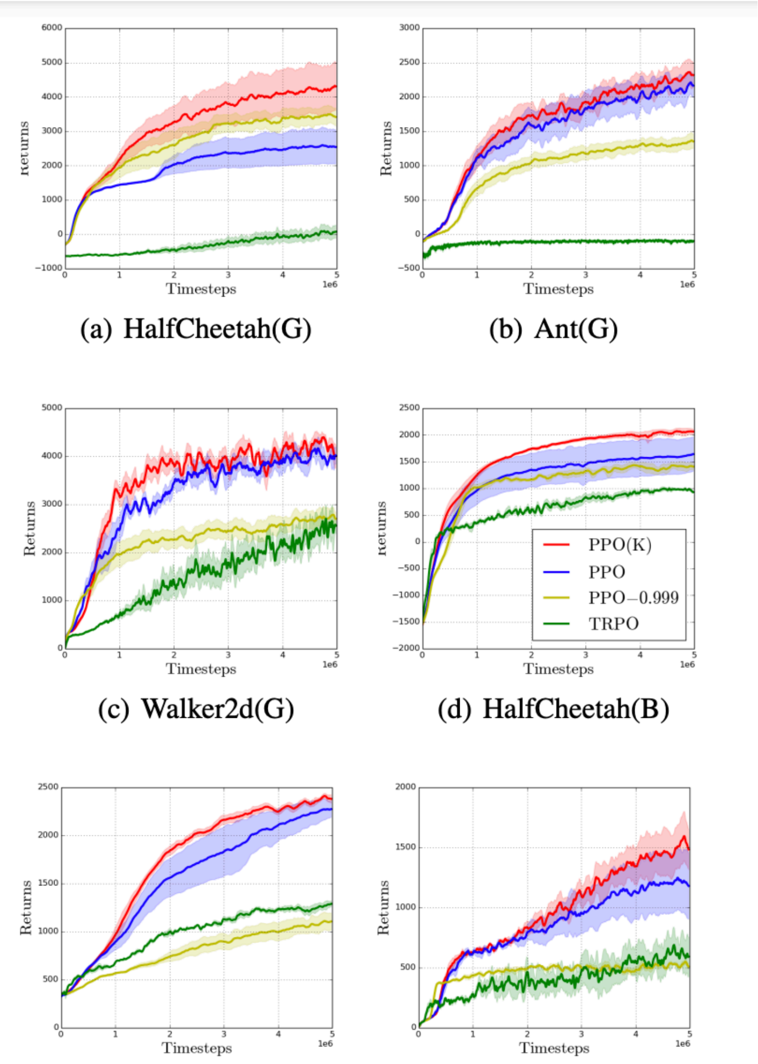
Implications for practical algorithms

- Insight: the practical heuristic can be seen as a partial gradient of the undiscounted objective
    - Some discrepancies: the horizon is truncated, so the problem is not Markovian…
- We can still improve current algorithms
    - Estimate advantage functions of a higher discount factors
    - Weigh the updates of PG algorithms

Experiments: advantage functions

Adapt Taylor expansions for
advantage estimates

$$V^{\pi}_{K,\gamma,\gamma'} := \sum_{k=0}^{K}((\gamma' - \gamma)(I - \gamma P^{\pi})^{-1} P^{\pi})^k V^{\pi}_{\gamma} \,.$$

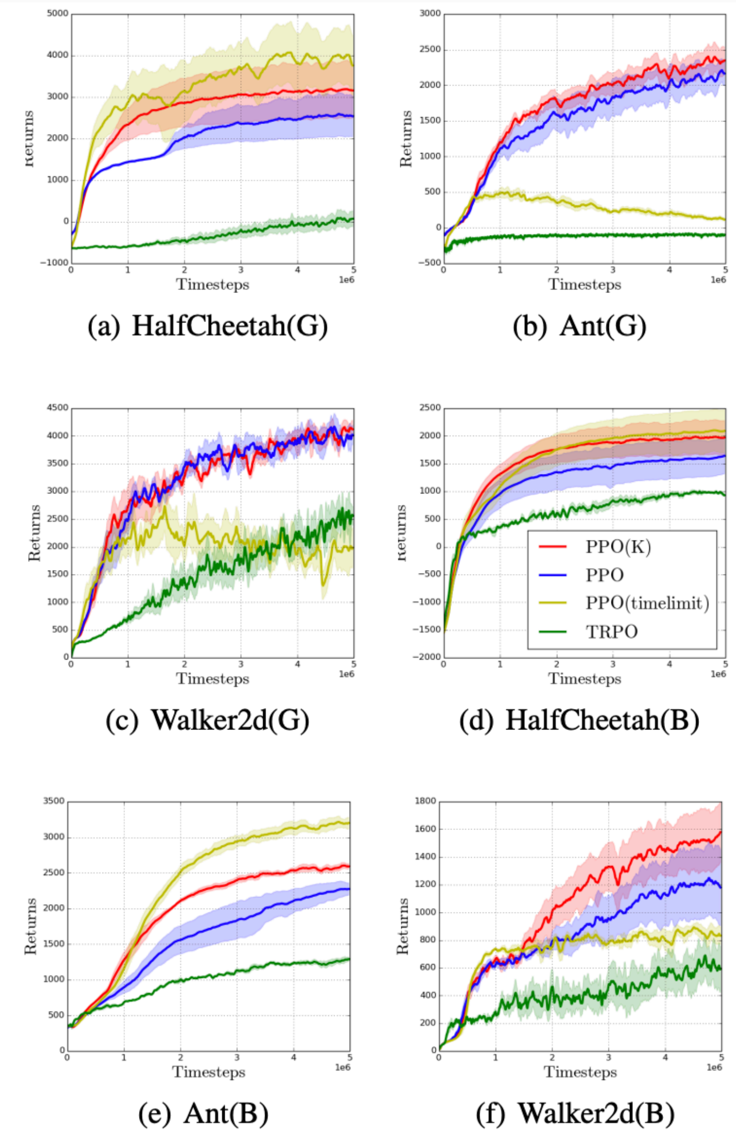


(a) HalfCheetah(G) (b) Ant(G) (c) Walker2d(G) (d) HalfCheetah(B) (e) Ant(B) (f) Walker2d(B)

Experiments: weighted updates

Weigh PG updates based on
K-th order expansion of the objective

$$\mathbb{E}_\pi\left[\sum_{t=0}^{\infty} w_{K,\gamma,\gamma'}(t)\, Q_t \nabla_\theta \log \pi_\theta(a_t|x_t) \,\middle|\, x_0 = x\right]$$



Summary

- **Theory**: discounted PG under γ ➡ Too conservative

$$E_\pi\left[\Sigma_{t=0}^{\infty}\gamma^t Q_\gamma^\pi(x_t,a_t)\nabla_\theta \log \pi(a_t|x_t)|x_0=x\right]$$

- **Theory**: discounted PG under γ' ➡ Too high variance

$$E_\pi\left[\Sigma_{t=0}^{\infty}(\gamma')^t Q_{\gamma'}^\pi(x_t,a_t)\nabla_\theta \log \pi(a_t|x_t)|x_0=x\right]$$

- **Practical heuristic**: can be derived as partial gradient ➡ Works in practice

$$E_\pi\left[\Sigma_{t=0}^{\infty}(\gamma')^t Q_\gamma^\pi(x_t,a_t)\nabla_\theta \log \pi(a_t|x_t)|x_0=x\right]$$