PAGE-PG: A Simple and Loopless Variance-Reduced Policy Gradient Method with Probabilistic Gradient Estimation

M. Gargiani*, A. Zanelli*, A. Martinelli*, T. H. Summers†, J. Lygeros*

*ETH Zurich, †University of Texas at Austin

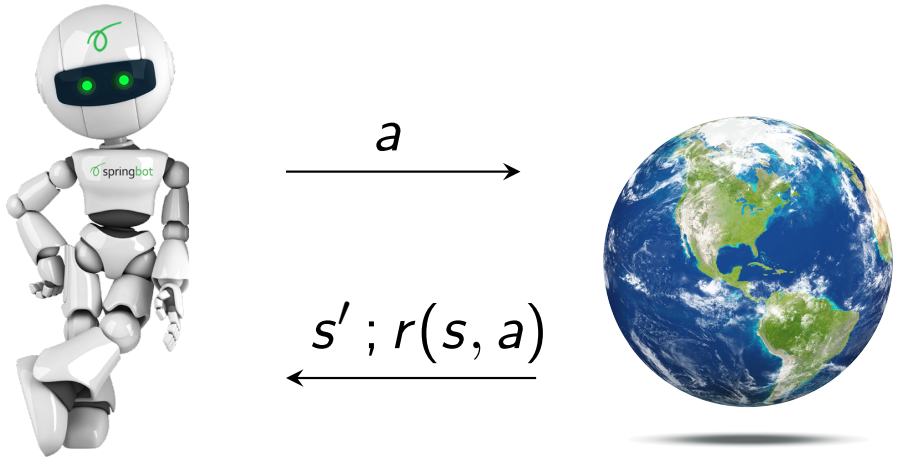
correspondence to: gmatilde@ethz.com

International Conference on Machine Learning (ICML)
July 17-23 2022

Overview

REINFORCE-Type Policy Gradient Methods

Problem Setting



- Markov Decision Processes (MDPs)
  $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho\}$
- stochastic stationary policies $\pi(\cdot \,|\, s)$
- trajectory $\tau = \{s_h, a_h\}_{h=0}^{H-1}$, where $s_0 \sim \rho$, $s_{h+1} \sim P(\cdot \,|\, s_h, a_h)$, $a_h \sim \pi(\cdot \,|\, s_h)$
- find a policy π^* that maximizes $V^\pi(\rho) = \mathbb{E}\left[\sum_{h=0}^{H-1} \gamma^h r(s_h, a_h) \,\middle|\, \pi\right]$
- episodic & discounted MDPs with bounded reward:
  $H < \infty$, $\gamma \in (0,1)$ and $r : \mathcal{S} \times \mathcal{A} \to [-U, U]$ with $U > 0$

REINFORCE-Type Policy Gradient Methods

Problem Setting

- tabular setting: $|\mathcal{S}| \times |\mathcal{A}| \rightarrow$ *curse-of-dimensionality*
- parametric policies $\Pi_\theta = \left\{\pi_\theta \,|\, \theta \in \mathbb{R}^d\right\}$ with $d << |\mathcal{S}| \times |\mathcal{A}|$

We aim at solving the following problem

$$\max_{\theta \in \mathbb{R}^d} V^{\pi_\theta}(\rho)$$

Policy Gradient Theorem

$$\nabla_\theta V^{\pi_\theta}(\rho) = \mathbb{E}_{\tau \sim p(\tau \,|\, \theta)} \left[\sum_{h=0}^{H-1} \nabla_\theta \log \pi_\theta(a_h \,|\, s_h) R(\tau) \right] .$$

Gradient Ascent

$$\theta \leftarrow \theta + \eta \nabla_\theta V^{\pi_\theta}(\rho), \quad \eta > 0$$

REINFORCE-Type Policy Gradient Methods

- in model-free RL, we can only estimate the true gradient
- given an unbiased estimator $g(\tau_i \,|\, \theta)$ and a collection of N trajectories with $\tau_i \sim p(\tau \,|\, \theta)$, we update the parameter as follow

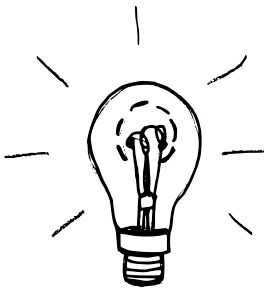
$$\theta \leftarrow \theta + \eta \frac{1}{N} \sum_{i=1}^{N} g(\tau_i \,|\, \theta)$$

- **REINFORCE**
  $g^{\mathrm{RF}}(\tau_i \,|\, \theta) = \sum_{h=0}^{H-1} \nabla_\theta \log \pi_\theta(a_h^i \,|\, s_h^i) R(\tau_i)$
- **GPOMDP**
  $g^{\mathrm{GPOMDP}}(\tau_i \,|\, \theta) = \sum_{h=0}^{H-1} \nabla_\theta \log \pi_\theta(a_h^i \,|\, s_h^i) \gamma^h \tilde{R}(\tau_i^h)$

these estimators are affected by high-variance $\rightarrow$ unsatisfactory sample-complexity, i.e. $\mathcal{O}(\epsilon^{-4})$

trajectory distribution depends on $\theta \rightarrow$ generate more samples is a naive solution!

Variance-Reduction For Policy Gradient



We could deploy variance-reduction techniques!

- successfully used in supervised learning [Allen-Zhu, 2017], [Nguyen et al., 2017]

Key ingredients & recent trends:

- double-loop structure: alternating between full gradient computation and updates based on interpolation of the full gradient with small-batch estimators
- in recent works the double-loop structure tends to be avoided: momentum-based methods [Cutkosky & Orabona, 2019], or probabilistic switch update [Kovalev et al., 2020], [Li et al., 2021]

From finite-sum problems to RL setting: [Papini et al., 2018]

- "infinite" dataset $\leftarrow$ large-batch estimator
- distribution-shift $\leftarrow$ importance-sampling

Variance-Reduction For Policy Gradient

The main variance-reduced REINFORCE-type PG methods

- **SVRPG** [Papini et al., 2018]

  double-loop structure
  sample complexity $\mathcal{O}\left(\epsilon^{-10/3}\right)$

- **STORM-PG** [Yuan et al., 2020]

  loopless structure ✔
  momentum-based update
  sample complexity $\mathcal{O}\left(\epsilon^{-3}\right)$

- **SRVRPG** [Xu et al., 2021]

  double-loop structure
  sample complexity $\mathcal{O}\left(\epsilon^{-3}\right)$

- **IS-MBPG** [Huang et al., 2021]

  loopless structure ✔
  momentum-based update
  sample complexity $\mathcal{O}\left(\epsilon^{-3}\right)$

(*) Check out our paper for more details!

PAGE-PG: Probabilistic Gradient Estimation for Policy Gradient

We propose a novel variance-reduced REINFORCE-type PG method

- inspired by PAGE estimator for finite-sum problems [Li et al., 2021]
- probabilistic switch between two estimators
- $N >> B$ (large and small batch-sizes) and $\mathbb{E}\left[g^{\omega_{\theta_t}}(\tau_i \,|\, \theta_{t-1})\right] = \mathbb{E}\left[g(\tau_i \,|\, \theta_{t-1})\right]$ (importance-sampling)
- $\theta_{t+1} = \theta_t + \eta v_t$ where

$$v_t = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} g(\tau_i \,|\, \theta_t) & \text{with } p_t \\ \frac{1}{B}\sum_{i=1}^{N} g(\tau_i \,|\, \theta_t) + v_{t-1} - \frac{1}{B}\sum_{i=1}^{B} g^{\omega_{\theta_t}}(\tau_i \,|\, \theta_{t-1}) & \text{with } 1 - p_t \end{cases}$$

- loopless structure ✔

Probability of switch p_t

p_t determines the frequency with which the gradient estimate based on a large batch is updated. $p_t = p = 1 \rightarrow$ REINFORCE/GPOMDP.

PAGE-PG: ProbAbilistic Gradient Estimation for Policy Gradient

Theoretical Analysis

- we consider typical assumptions for comparison reasons & $p_t = p$
- $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla_\theta V(\theta_t)\|^2\right] \leq \frac{2(V^*-V(\theta_0))}{\eta T} + \frac{\sigma^2}{N} + \frac{\sigma^2}{pNT} \implies$
  $\mathcal{O}\left(\frac{1}{T}+\frac{1}{N}+\frac{1}{NT}\right)$ & $\mathcal{O}\left(\frac{1}{T}+\frac{1}{N}\right)$ for $p = B/N$ and $B = \mathcal{O}(1)$
- sample complexity $\mathcal{O}(\epsilon^{-3})$

Empirical Results

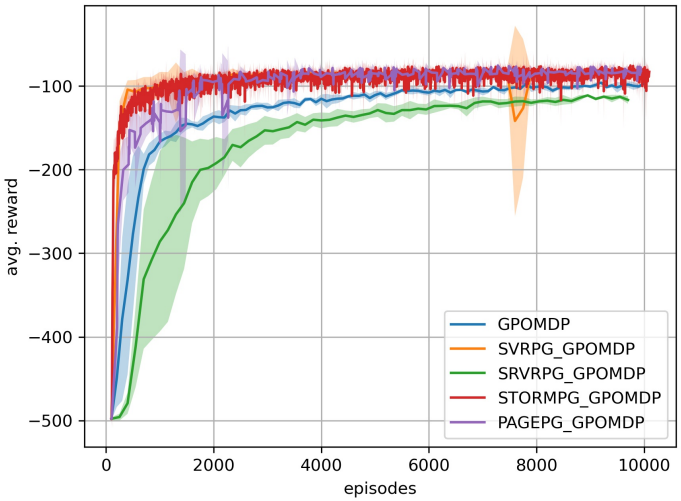


Figure: Acrobot

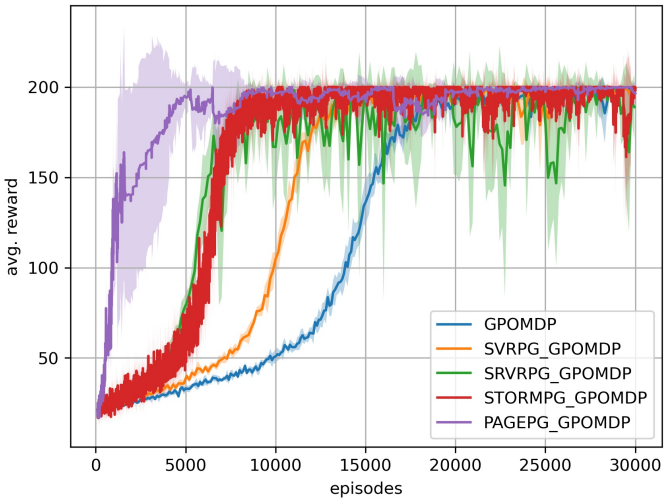


Figure: Cartpole

Conclusions & Promising Future Extensions

Conclusions

- novel variance-reduced REINFORCE-type PG method
- loopless structure
- probabilistic switch
- sample complexity matches its most competitive counterparts
- competitive empirical performance

Future Work

- assumption on bounded variance of importance sampling weight is unrealistic [Zhang et al., 2021] $\leftarrow$ trust-region approach
- exploiting the variance of the gradient estimator for exploration vs exploitation [Chung et al., 2021] $\leftarrow$ iteration-varying strategies for adjusting p_t

Thanks :)