



REUSING TRAJECTORIES IN POLICY GRADIENTS ENABLES FAST CONVERGENCE

A. MONTENEGRO, F. MANSUTTI, M. MUSSI, M. PAPINI, AND A. M. METELLI
alessandro.montenegro@polimi.it



CHALLENGE: REUSE HISTORICAL DATA

POLICY GRADIENTS ARE SAMPLE INEFFICIENT
They require $\mathcal{O}(\epsilon^{-2})$ trajectories to converge to an ϵ -approximate stationary point.

THE PROBLEM

The problem is intrinsic in the **estimator's variance**. Vanilla PG estimators (REINFORCE or GPOMDP) at each iteration:

- (i) **collect** a batch of N **fresh on-policy trajectories**
- (ii) **compute** the gradient and **discard** the trajectories

DATA REUSE

The literature focused on **gradient reuse**, lowering the sample complexity to $\mathcal{O}(\epsilon^{-1.5})$. **Reusing trajectories** is more **natural**, but it is still **unexplored** in theory.

TECHNICAL CHALLENGE

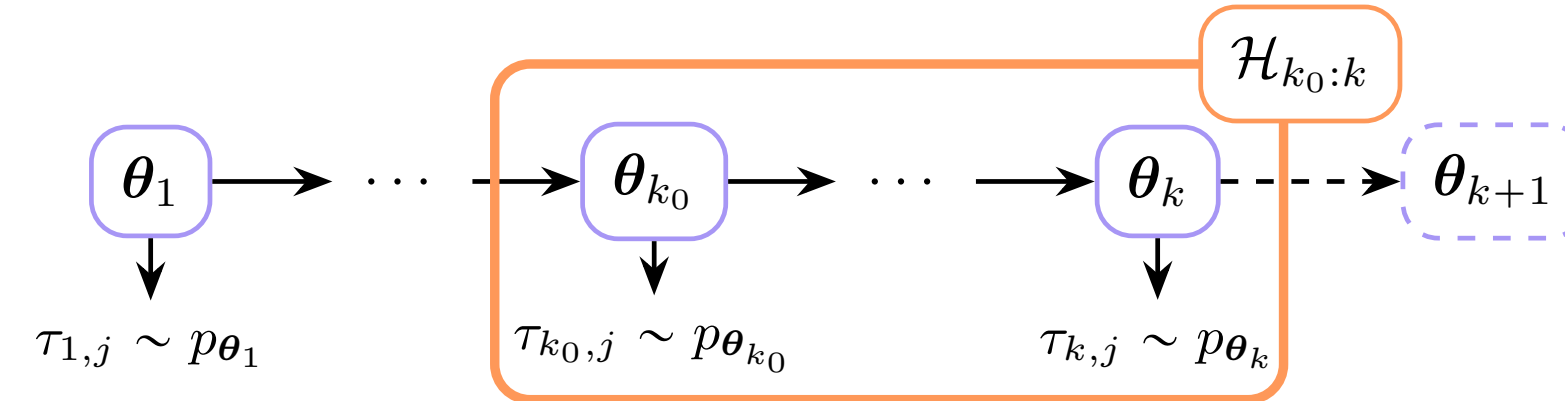
Reusing past trajectories makes the **gradient estimator biased**. Indeed, the **past trajectories** have been used to **generate the target policy**.

REUSING DATA FROM THE LEARNING PROCESS: TARGET BIAS

SLIDING WINDOW

Reuse data from the last $\omega_k := \min\{k, \omega\}$ iterations

$k_0 := k - \omega_k + 1$ is the starting index



HISTORY AT ITERATION k

$$\mathcal{H}_k = \{\theta_1, \{\tau_{1,j}\}_{j=1}^N, \dots, \theta_k, \{\tau_{k,j}\}_{j=1}^N\}$$

$\tau_{i,j} \sim p_{\theta_i}$ sampled independently

LEARNING PROCESS

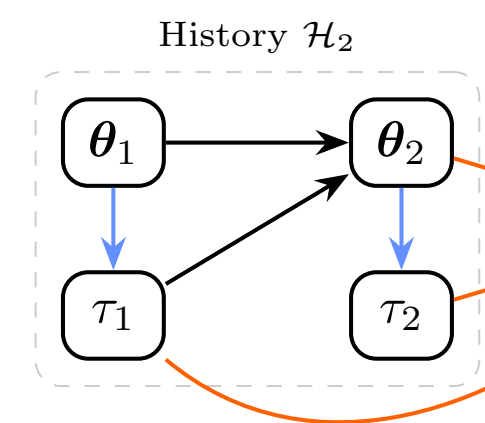
$$\theta_{k+1} \leftarrow \theta_k + \zeta_k \mathbf{G}(\theta_k, \mathcal{H}_{k_0:k})$$

$\mathbf{G}(\theta_k, \mathcal{H}_{k_0:k})$: gradient estimator in θ_k exploiting $\mathcal{H}_{k_0:k}$

CHALLENGES

- ✗ $\mathbf{G}(\mathcal{H}_{k_0:k})$ suffers from the **target bias**
- ✗ $\hat{\nabla}^{\text{MIW}} J(\theta)$ variance of IWs may explode + need to select β_i

TARGET BIAS: AN EXAMPLE



Let $\beta_1(\tau_1) = \beta_2(\tau_2) = 1/2$

$$\mathbb{E}[\hat{\nabla}_2^{\text{MIW}} J(\theta_2) | \theta_2] \neq \nabla J(\theta_2)$$

INCORPORATING OFF-POLICY DATA

$$J(\theta) = \mathbb{E}_{\tau \sim p_\theta} [R(\tau)] \leftarrow \text{objective}$$

$\mathbf{g}_\theta(\tau) \leftarrow$ single-trajectory gradient estimator

ON-POLICY

$$\hat{\nabla} J(\theta) = \frac{1}{N} \sum_{j=1}^N \mathbf{g}_\theta(\tau_j)$$

$\tau \sim p_\theta$ for every $\tau \in \{\tau_j\}_{j=1}^N$

SINGLE IMPORTANCE WEIGHTING (SIW)

$$\hat{\nabla}^{\text{SIW}} J(\theta) = \frac{1}{N} \sum_{j=1}^N \frac{p_\theta(\tau_j)}{p_{\theta_b}(\tau_j)} \mathbf{g}_\theta(\tau_j)$$

$\tau \sim p_{\theta_b}$ for every $\tau \in \{\tau_j\}_{j=1}^N$

MULTIPLE IW

$$\hat{\nabla}_m^{\text{MIW}} J(\theta) = \sum_{i=1}^m \frac{1}{N_i} \sum_{j=1}^{N_i} \beta_i(\tau_{i,j}) \frac{p_\theta(\tau_{i,j})}{p_{\theta_i}(\tau_{i,j})} \mathbf{g}_\theta(\tau_{i,j})$$

$\tau \sim p_{\theta_i}$ for every $\tau \in \{\tau_{i,j}\}_{j=1}^{N_i}$ | $\sum_{i=1}^m \beta_i(\tau) = 1$ with $\beta_i(\tau) \geq 0$ $\forall \tau \in \mathcal{T}$ and $\forall i \in [m]$

RT-PG: REUSING TRAJECTORIES – POLICY GRADIENT

MULTIPLE POWER MEAN (MPM) ESTIMATOR

$$\hat{\nabla}_{\omega_k}^{\text{MPM}} J(\theta_k) = \frac{1}{N} \sum_{i=k_0}^k \sum_{j=1}^N \frac{\alpha_{i,k} p_{\theta_k}(\tau_{i,j}) \mathbf{g}_{\theta_k}(\tau_{i,j})}{(1 - \lambda_{i,k}) p_{\theta_i}(\tau_{i,j}) + \lambda_{i,k} p_{\theta_k}(\tau_{i,j})}$$

- (A) $\sum_{i=k_0}^k \alpha_{i,k} = 1$ with $\alpha_{i,k} \geq 0$
- (B) $\lambda_{i,k} \in [0, 1]$
- (C) Defensive when $\lambda_{i,k} > 0$
- (D) No need for storing past θ s

ASSUMPTIONS AND INDUCED PROPERTIES

- (A₁) $\log \pi_\theta$ REGULARITY
- (A₂) $\sup_{\theta_1, \theta_2 \in \Theta} \chi^2(p_{\theta_1} \| p_{\theta_2}) \leq D - 1$
- (P₁) SMOOTHNESS OF $J(\theta)$
- (P₂) $\sup_{\theta, \tau \in \Theta \times \mathcal{T}} \|\mathbf{g}_\theta(\tau)\|_2 \leq G_1$

RESULTS

Theorem. Let $\bar{\theta}$ be independent of $\mathcal{H}_k$. For any $\delta \in (0, 1)$, for every $i \in [k_0, k]$ select

$$\lambda_{i,k} = \sqrt{\frac{4(d_\Theta \log 6 + \log \frac{1}{\delta})}{3DN\omega_k}} \quad \text{and} \quad \alpha_{i,k} = \frac{1}{\omega_k}.$$

Then w.p. $1 - \delta$:

$$\|\hat{\nabla}_{\omega_k}^{\text{MPM}} J(\bar{\theta}) - \nabla J(\bar{\theta})\|_2 \leq 8G_1 \sqrt{\frac{D(d_\Theta \log 6 + \log \frac{1}{\delta})}{N\omega_k}}$$

Theorem. Let $\mathcal{H}_k$ be generated by RT-PG. For any $\delta \in (0, 1)$, for every $i \in [k_0, k]$ select

$$\lambda_{i,k} = \tilde{\mathcal{O}} \left(\sqrt{\frac{d_\Theta \log \frac{1}{\delta}}{DN\omega_k}} \right) \quad \text{and} \quad \alpha_{i,k} = \frac{1}{\omega_k}.$$

Then w.p. $1 - \delta$:

$$\|\hat{\nabla}_{\omega_k}^{\text{MPM}} J(\theta_k) - \nabla J(\theta_k)\|_2 \leq \tilde{\mathcal{O}} \left(G_1 \sqrt{\frac{D(d_\Theta + \log \frac{1}{\delta})}{N\omega_k}} \right)$$

Covering argument over Θ

SAMPLE COMPLEXITY

PARTIAL REUSE

$$\omega < K$$

$$NK \geq \tilde{\mathcal{O}} \left(\frac{G_1^2 d_\Theta D}{\epsilon^2 \omega} \right)$$

FULL REUSE

$$\omega \geq K$$

$$NK \geq \tilde{\mathcal{O}} \left(\frac{G_1^2 d_\Theta D}{\epsilon} \right)$$

for an ϵ -approximate stationary point

✓ Rate adaptive w.r.t. ω

✓ No storage of past policies

✗ Stores up to $\tilde{\mathcal{O}}(\epsilon^{-1})$ trajectories

✓ Step size independent of ϵ

✓ $\tilde{\mathcal{O}}(\epsilon^{-1})$ under full reuse

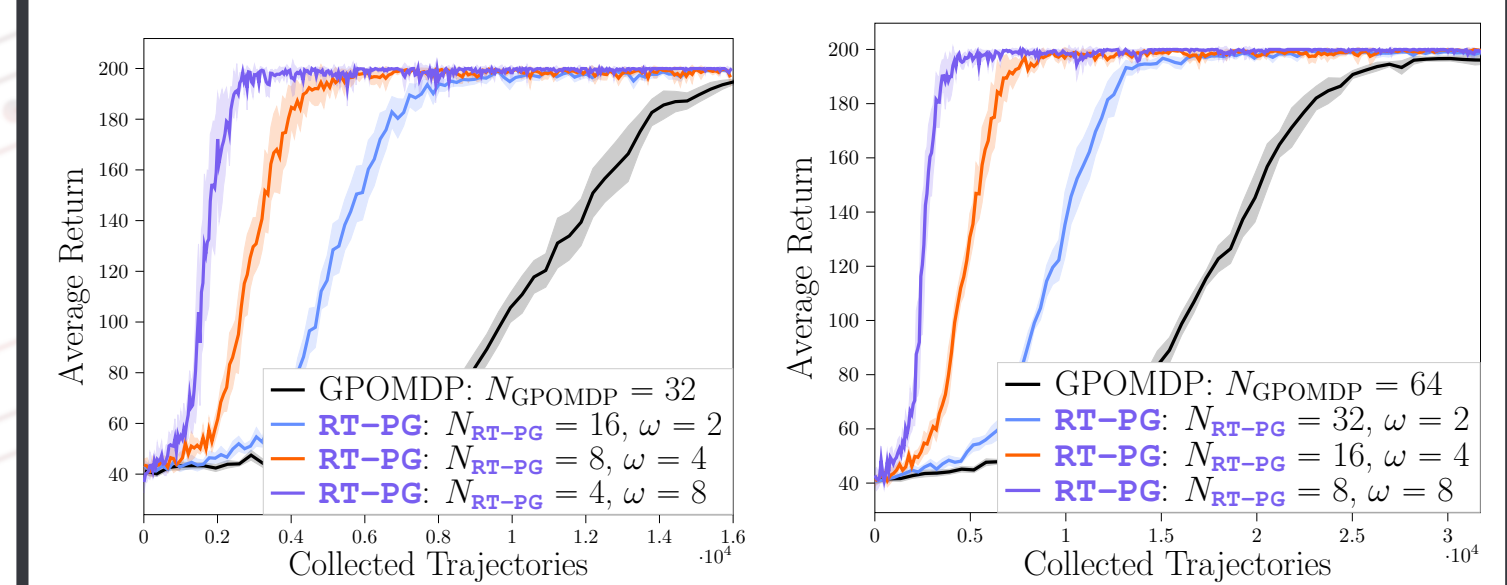
✗ Relies on D and d_Θ

DOES IT WORK?

- ✓ Limited $N\omega_k$ values
- ✓ Adaptive $\lambda_{i,k}$ and $\alpha_{i,k}$ coefficients

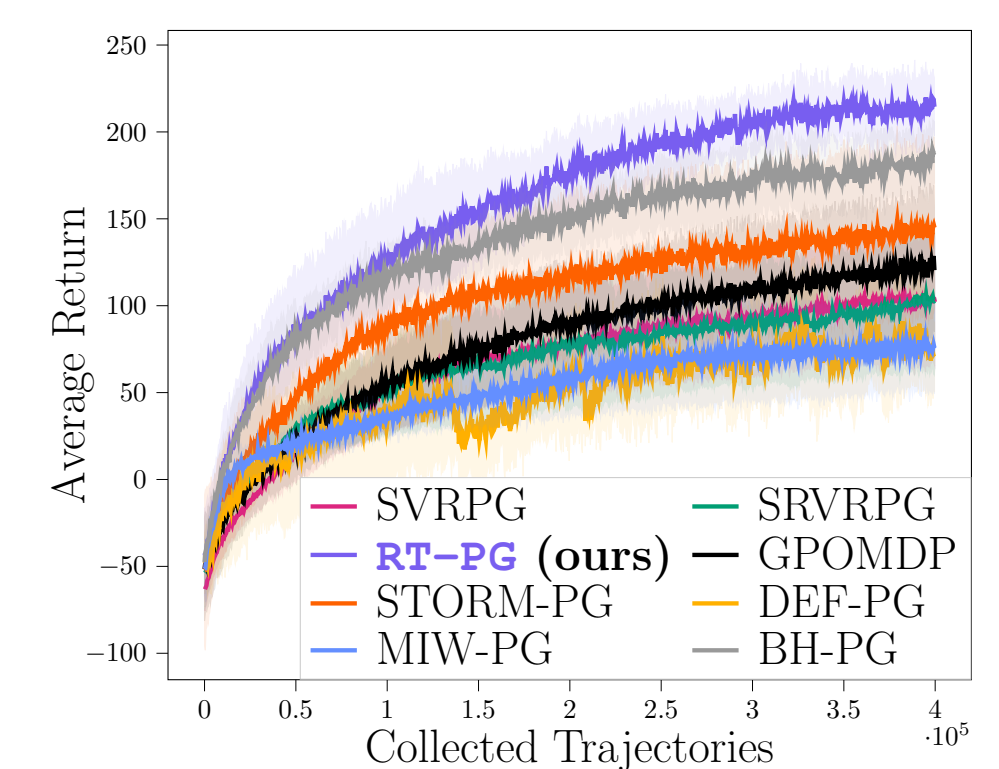
TO REUSE OR NOT TO REUSE?

CART POLE



REUSING TRAJECTORIES VS. GRADIENTS

HALF CHEETAH



REFERENCES

- A. M. Metelli, M. Papini, F. Faccio, and M. Restelli. Policy optimization via importance sampling. In *NeurIPS*, 2018.
M. Papini, D. Binaghi, G. Canonaco, M. Pirota, and M. Restelli. Stochastic variance-reduced policy gradient. In *ICML*, 2018.