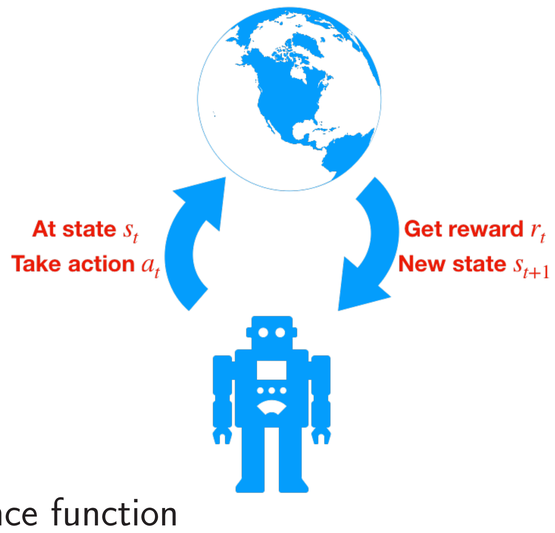
# Sample Efficient Policy Gradient Methods with Recursive Variance Reduction

**Pan Xu** and **Felicia Gao** and **Quanquan Gu**
University of California, Los Angeles

## Policy Gradient

- **Markov decision process (MDP)**
  - State $s \in \mathcal{S}$
  - Action $a \in \mathcal{A}$
  - Transition transition $P(s'|s,a)$
  - Reward $r(s,a)$
  - Policy $\pi_{\boldsymbol{\theta}}(a|s)$




- **Goal:** maximize the performance function

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta})}\Big[\sum_{h=0}^{H-1}\gamma^h r(s_h,a_h)\Big]$$

  - **Trajectory:** $\tau = (s_0, a_0, s_1, r_1, \ldots, s_{H-1}, a_{H-1}, s_H)$
  - **Trajectory distribution:** $p(\tau|\boldsymbol{\theta}) = \rho(s_0)\prod_{h=0}^{H-1}\pi_{\boldsymbol{\theta}}(a_h|s_h)P(s_{h+1}|s_h,a_h)$
- **GPOMDP policy gradient estimator:**

$$g(\tau_i|\boldsymbol{\theta}) = \sum_{h=0}^{H-1}\gamma^h r(s_h^i,a_h^i)\sum_{t=0}^{h}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a_t^i|s_t^i) \approx \nabla J(\boldsymbol{\theta})$$

## SRVR-PG

- **Input**: number of epochs $S$, epoch length $m$, step size $\eta$, batch size $N$, mini-batch size $B$, gradient estimator $g(\tau|\boldsymbol{\theta})$, initial parameter $\widetilde{\boldsymbol{\theta}}^0 = \boldsymbol{\theta}_0$

**for** $s = 0, \ldots, S-1$ **do**
  $\boldsymbol{\theta}_0^{s+1} = \widetilde{\boldsymbol{\theta}}^s$
  sample $N$ trajectories $\{\tau_i\}$
  $\mathbf{v}_0^{s+1} = \frac{1}{N}\sum_{i=1}^{N} g(\tau_i|\boldsymbol{\theta}_0^{s+1})$
  $\boldsymbol{\theta}_1^{s+1} = \boldsymbol{\theta}_0^{s+1} + \eta\mathbf{v}_0^{s+1}$
  **for** $t = 1, \ldots, m-1$ **do**
    sample $B$ trajectories $\{\tau_j\}$
    $\mathbf{v}_t^{s+1} = \mathbf{v}_{t-1}^{s+1} + \frac{1}{B}\sum_{j=1}^{B}(g(\tau_j|\boldsymbol{\theta}_t^{s+1}) - g_\omega(\tau_j|\boldsymbol{\theta}_{t-1}^{s+1}))$
    $\boldsymbol{\theta}_{t+1}^{s+1} = \boldsymbol{\theta}_t^{s+1} + \eta\mathbf{v}_t^{s+1}$
  **end for**
**end for**
**output** $\boldsymbol{\theta}_{\text{out}}$ uniformly from $\{\boldsymbol{\theta}_t^s\}$

## Step-wise Importance Sampling

- **Policy gradient estimator**

$$g_\omega(\tau|\boldsymbol{\theta}_{t-1}^{s+1}) = \sum_{h=0}^{H-1}\omega_{0:h}(\tau)\Big[\sum_{t=0}^{h}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}_{t-1}^{s+1}}(a_t|s_t)\Big]\gamma^h r(s_h,a_h)$$

- **Importance weight**

$$\omega_{0:h}(\tau) := \omega_{0:h}(\tau|\boldsymbol{\theta}_{t-1}^{s+1},\boldsymbol{\theta}_t^{s+1}) = \prod_{k=0}^{h}\frac{\pi_{\boldsymbol{\theta}_{t-1}^{s+1}}(a_k|s_k)}{\pi_{\boldsymbol{\theta}_t^{s+1}}(a_k|s_k)}$$

## Convergence Analysis

- **Assumptions**
  - **Smoothness:** for all $\boldsymbol{\theta}, s, a$, the policy satisfies

$$\|\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)\| \le G, \quad \big\|\nabla^2_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)\big\|_2 \le M$$

  - **Bounded variance:** there exists a constant $\xi > 0$ such that $\mathrm{Var}\big(g(\tau|\boldsymbol{\theta})\big) \le \xi^2$, for all policy $\pi_{\boldsymbol{\theta}}$
  - **Importance weight:** there is a constant $W < \infty$ such that $\omega(\cdot|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = p(\cdot|\boldsymbol{\theta}_1)/p(\cdot|\boldsymbol{\theta}_2)$ satisfies $\mathrm{Var}(\omega(\tau|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)) \le W, \quad \forall\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \in \mathbb{R}^d, \tau\sim p(\cdot|\boldsymbol{\theta}_2)$
- **Convergence of SRVR-PG**

Let step size $\eta \le 1/(4L)$ and batch size $B \ge Cm\eta G^4 MW\gamma(1-\gamma)^{-3}$

$$\mathbb{E}\big[\big\|\nabla J\big(\boldsymbol{\theta}_{\text{out}}\big)\big\|_2^2\big] \le \frac{8(J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0))}{\eta Sm} + \frac{6\xi^2}{N}$$

**Remark:** independent of $H$ due to the step-wise importance weight

## Comparison on Sample Complexity

- **Sample complexity**
  toatal number of trajectories needed to achieve $\|\nabla J(\boldsymbol{\theta})\|_2^2 \le \epsilon$
- Setting $N = O(1/\epsilon)$, $B = m = S = O(1/\epsilon^{1/2})$, we obtain the sample complexity of SRVR-PG: $O(1/\epsilon^{3/2})$

| Algorithms | Complexity |
|---|---|
| REINFORCE (Williams, 1992) | $O(1/\epsilon^2)$ |
| PGT (Sutton et al., 2000) | $O(1/\epsilon^2)$ |
| GPOMDP (Baxter and Bartlett, 2001) | $O(1/\epsilon^2)$ |
| SVRPG (Papini et al., 2018) | $O(1/\epsilon^2)$ |
| SVRPG (Xu et al., 2019) | $O(1/\epsilon^{5/3})$ |
| SRVR-PG | $O(1/\epsilon^{3/2})$ |

## Experiments

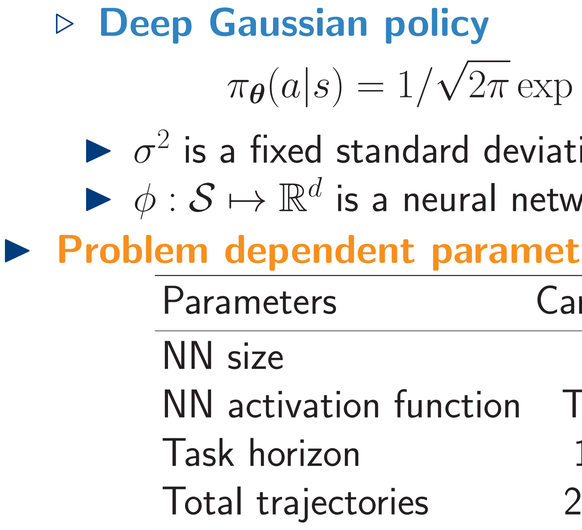
- **Setup**
  - **Environments:** Cartpole, Mountain Car, Pendulum
  - **Deep Gaussian policy**

$$\pi_{\boldsymbol{\theta}}(a|s) = 1/\sqrt{2\pi}\exp\big(-(f(\boldsymbol{\theta};\phi(s)) - a)^2/(2\sigma^2)\big)$$

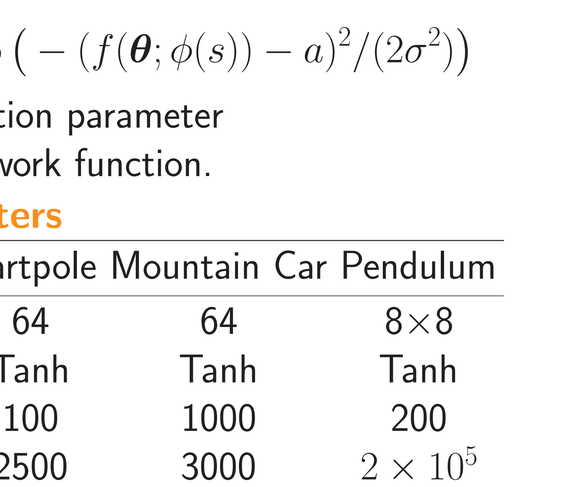
    - $\sigma^2$ is a fixed standard deviation parameter
    - $\phi : \mathcal{S} \mapsto \mathbb{R}^d$ is a neural network function.
- **Problem dependent parameters**

| Parameters | Cartpole | Mountain Car | Pendulum |
|---|---|---|---|
| NN size | 64 | 64 | 8×8 |
| NN activation function | Tanh | Tanh | Tanh |
| Task horizon | 100 | 1000 | 200 |
| Total trajectories | 2500 | 3000 | $2\times10^5$ |

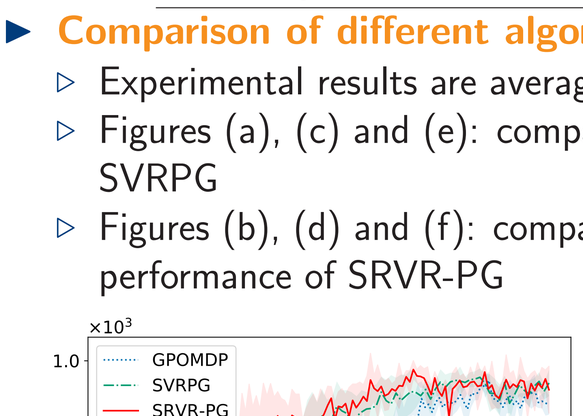
- **Comparison of different algorithms**
  - Experimental results are averaged over 10 repetitions
  - Figures (a), (c) and (e): comparison with baselines GPOMDP and SVRPG
  - Figures (b), (d) and (f): comparison of different batch size $B$ on the performance of SRVR-PG



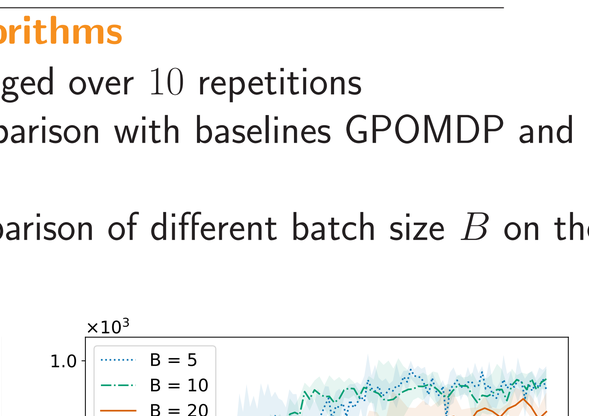

(a) Cartpole



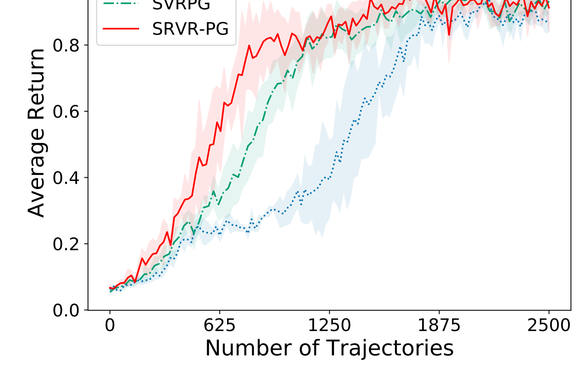

(b) Cartpole



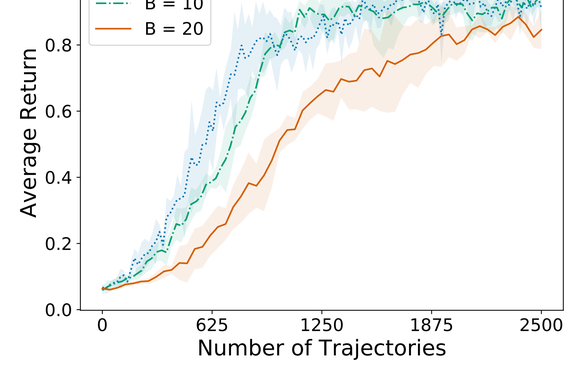

(c) Mountain Car




(d) Mountain Car




(e) Pendulum




(f) Pendulum