Variance Reduced Experience Replay for Policy Optimization with Partial Trajectory Reuse

Hua Zheng[1] Wei Xie[1]

[1]Department of Mechanical and Industrial Engineering
Northeastern University

Motivation

- RL is known for its sample inefficiency:

Motivation

- RL is known for its sample inefficiency:
    - "learning from limited interaction remains a key challenge" (Schwarzer et al. 2020)

Motivation

- RL is known for its sample inefficiency:
  - "learning from limited interaction remains a key challenge" (Schwarzer et al. 2020)
  - "Learning on the real system from limited samples" is listed as one of 7 major challenge (Dulac-Arnold et al., 2021)

Motivation

- RL is known for its sample inefficiency:
  - "learning from limited interaction remains a key challenge" (Schwarzer et al. 2020)
  - "Learning on the real system from limited samples" is listed as one of 7 major challenge (Dulac-Arnold et al., 2021)
- "On-policy learning, in simplest form, discard incoming data immediately, after a single update." (Schaul et al., 2015)

Motivation

- RL is known for its sample inefficiency:
  - "learning from limited interaction remains a key challenge" (Schwarzer et al. 2020)
  - "Learning on the real system from limited samples" is listed as one of 7 major challenge (Dulac-Arnold et al., 2021)
- "On-policy learning, in simplest form, discard incoming data immediately, after a single update." (Schaul et al., 2015)
  - strongly correlated updates breaks the i.i.d. assumption of SGD

Motivation

- RL is known for its sample inefficiency:
  - "learning from limited interaction remains a key challenge" (Schwarzer et al. 2020)
  - "Learning on the real system from limited samples" is listed as one of 7 major challenge (Dulac-Arnold et al., 2021)
- "On-policy learning, in simplest form, discard incoming data immediately, after a single update." (Schaul et al., 2015)
  - strongly correlated updates breaks the i.i.d. assumption of SGD
  - the rapid forgetting of possibly rare experiences.

Motivation

- RL is known for its sample inefficiency:
  - "learning from limited interaction remains a key challenge" (Schwarzer et al. 2020)
  - "Learning on the real system from limited samples" is listed as one of 7 major challenge (Dulac-Arnold et al., 2021)
- "On-policy learning, in simplest form, discard incoming data immediately, after a single update." (Schaul et al., 2015)
  - strongly correlated updates breaks the i.i.d. assumption of SGD
  - the rapid forgetting of possibly rare experiences.
  - **Idea**: experience replay (reusing historical samples) and off-policy learning.

Motivation

- RL is known for its sample inefficiency:
  - "learning from limited interaction remains a key challenge" (Schwarzer et al. 2020)
  - "Learning on the real system from limited samples" is listed as one of 7 major challenge (Dulac-Arnold et al., 2021)
- "On-policy learning, in simplest form, discard incoming data immediately, after a single update." (Schaul et al., 2015)
  - strongly correlated updates breaks the i.i.d. assumption of SGD
  - the rapid forgetting of possibly rare experiences.
  - **Idea**: experience replay (reusing historical samples) and off-policy learning.
  - **What is the problem?**

Motivation

- RL is known for its sample inefficiency:
    - "learning from limited interaction remains a key challenge" (Schwarzer et al. 2020)
    - "Learning on the real system from limited samples" is listed as one of 7 major challenge (Dulac-Arnold et al., 2021)
- "On-policy learning, in simplest form, discard incoming data immediately, after a single update." (Schaul et al., 2015)
    - strongly correlated updates breaks the i.i.d. assumption of SGD
    - the rapid forgetting of possibly rare experiences.
    - **Idea**: experience replay (reusing historical samples) and off-policy learning.
    - **What is the problem?**
        - how to avoid high variance in the policy gradient (Metelli et al., 2020; Schlegel et al., 2019; Zheng et al., 2021)
        - "how prioritizing which transitions are replayed" (Schaul et al., 2015)

Inflated Variance in Off-policy Policy Optimization

To reuse the historical samples, RL needs to perform the **importance sampling** (IS) on full trajectories to adjust the distributional mismatch between the target and sampling/behavior policies.

Inflated Variance in Off-policy Policy Optimization

To reuse the historical samples, RL needs to perform the **importance sampling** (IS) on full trajectories to adjust the distributional mismatch between the target and sampling/behavior policies.

Problem of IS: The importance weights (or likelihood ratios) are

- the products of policy ratios for all transitions within a trajectory (Metelli et al., 2020; Zheng et al., 2021).

Inflated Variance in Off-policy Policy Optimization

To reuse the historical samples, RL needs to perform the **importance sampling** (IS) on full trajectories to adjust the distributional mismatch between the target and sampling/behavior policies.

Problem of IS: The importance weights (or likelihood ratios) are

- the products of policy ratios for all transitions within a trajectory (Metelli et al., 2020; Zheng et al., 2021).
- can have **high** or even **infinite** variance. (Andradóttir et al., 1995; Schlegel et al., 2019)

Inflated Variance in Off-policy Policy Optimization

To reuse the historical samples, RL needs to perform the **importance sampling** (IS) on full trajectories to adjust the distributional mismatch between the target and sampling/behavior policies.

Problem of IS: The importance weights (or likelihood ratios) are

- the products of policy ratios for all transitions within a trajectory (Metelli et al., 2020; Zheng et al., 2021).
- can have **high** or even **infinite** variance. (Andradóttir et al., 1995; Schlegel et al., 2019)
- As a result, importance sampling / likelihood ratio based policy gradient estimator inevitably suffers from high variance.

Proposed Approach

Motivated by the problems discussed above, we invented a new experience replay technique called **variance reduced experience replay (VRER)** and investigated its applicability to step–based policy optimization. It

- prioritizes the transitions that can reduce policy gradient variance.
- automatically selects historical transitions based on a comparison of gradient variance between historical transitions and current transitions.
- has theoretically and empirically shown that the MLR based policy gradient estimator improves sample efficiency and has superior performance in convergence.

Review: Episode–based versus Step–based

- **Episode–based approaches** are also known as Monte Carlo approaches: REINFORCE (Williams, 1992)
- **Step–based approaches** (also known as per-decision approaches): trust region policy optimization (TRPO) (Schulman et al., 2015), proximal policy optimization (PPO) (Schulman et al., 2017)

Problem Description: Infinite Horizon MDP

We formulate the problem of interest as infinite-horizon Markov decision process (MDP) specified by $(\mathcal{S}, \mathcal{A}, r, p, \boldsymbol{s}_1)$, where

- a transition dynamics distribution with conditional density $p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)$
- a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

The system start at an initial state $\boldsymbol{s}_1$ drawn from $p_1(\boldsymbol{s}_1)$. At time t,

- the agent observes the state $\boldsymbol{s}_t \in \mathcal{S}$ and takes an action $\boldsymbol{a}_t \in \mathcal{A}$ from a parametric policy $\pi(\boldsymbol{s}_t|\boldsymbol{a}_t; \boldsymbol{\theta})$ with parameter $\boldsymbol{\theta} \in \mathbb{R}^d$
- receives a reward $r_t(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \mathbb{R}$.

Problem Description: Infinite Horizon MDP

- **Return**: the total discounted reward from time-step t onwards. Defined as

$$r_t^\gamma = \sum_{t'=t}^{\infty} \gamma^{t'-t} r(\boldsymbol{s}_{t'}, \boldsymbol{a}_{t'})$$

  where $\gamma \in (0,1)$ denotes the discount factor.

- **Value Function**: state value functions $V^\pi(\boldsymbol{s})$ and the action function $Q^\pi(\boldsymbol{s},\boldsymbol{a})$ are defined to be the expected total discounted reward-to-go,

$$V^\pi(\boldsymbol{s}) = \mathbb{E}[r_1^\gamma|\boldsymbol{s}_1 = \boldsymbol{s};\pi] = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r(\boldsymbol{s}_t,\boldsymbol{a}_t)\middle| \boldsymbol{s}_1 = \boldsymbol{s};\pi\right] \quad (1)$$

$$Q^\pi(\boldsymbol{s},\boldsymbol{a}) = \mathbb{E}[r_1^\gamma|\boldsymbol{s}_1 = \boldsymbol{s},\boldsymbol{a}_1 = \boldsymbol{a};\pi] = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r(\boldsymbol{s}_t,\boldsymbol{a}_t)\middle| \boldsymbol{s}_1 = \boldsymbol{s},\boldsymbol{a}_1 = \boldsymbol{a};\pi\right]. \quad (2)$$

- **Objective**: $J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{s}\sim d^\pi(\boldsymbol{s}),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})}[r(\boldsymbol{s},\boldsymbol{a})]$,

  where $\mathbb{E}_{\boldsymbol{s}\sim d^\pi(\boldsymbol{s}),\boldsymbol{a}\sim\pi(\boldsymbol{a}|\boldsymbol{s}))}[\cdot]$ denotes the expected value with respect to
  - stationary state distribution $d^\pi(\boldsymbol{s}) = \int_{\mathcal{S}} \sum_{t=1}^{\infty} \gamma^{t-1} p(\boldsymbol{s}_1) p(\boldsymbol{s}_t = \boldsymbol{s}|\boldsymbol{s}_1;\pi) d\boldsymbol{s}_1$
  - policy distribution $\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})$.

Problem Description: Policy Optimization

- Under some regularity conditions, *Policy Gradient Theorem* (Sutton et al., 1999) reformulates the policy gradient as

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{s}\sim d^{\pi}(\boldsymbol{s}),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})}[\nabla \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s}) Q^{\pi}(\boldsymbol{s},\boldsymbol{a})] \quad (3)$$

- A widely used variation of (3) is to subtract a state value function from the return to reduce the variance of gradient estimation while keeping the bias unchanged (Bhatnagar et al., 2009, Lemma 2):

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{s}\sim d^{\pi}(\boldsymbol{s}),\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})}[\nabla \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s}) A^{\pi}(\boldsymbol{s},\boldsymbol{a}))] \quad (4)$$

  The difference $A^{\pi}(\boldsymbol{s},\boldsymbol{a}) = Q^{\pi}(\boldsymbol{s},\boldsymbol{a}) - V^{\pi}(\boldsymbol{s})$ is called advantage.

- The advantage function can be also expressed

$$A^{\pi}(\boldsymbol{s},\boldsymbol{a}) = r(\boldsymbol{s},\boldsymbol{a}) + \gamma\, \mathbb{E}_{\boldsymbol{s}'\sim p(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})}[V^{\pi}(\boldsymbol{s}')] - V^{\pi}(\boldsymbol{s}). \quad (5)$$

  It can be estimated by the temporal difference (TD) error

$$\delta(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a}) + \gamma \hat{V}(\boldsymbol{s}') - \hat{V}(\boldsymbol{s}) \quad (6)$$

  is an unbiased estimate of $A^{\pi}(\boldsymbol{s},\boldsymbol{a})$ (Bhatnagar et al., 2009, Lemma 3). Here, $\hat{V}(\boldsymbol{s})$ is an unbiased estimate of value function at state $\boldsymbol{s}$.

Individual/Mixture Likelihood Ratio (ILR/MLR)

Let stationary probabilities of state-action pair to be $\rho_\theta(\boldsymbol{s}, \boldsymbol{a}) = \pi_\theta(\boldsymbol{a}|\boldsymbol{s}) d^\pi(\boldsymbol{s})$.

Individual Likelihood Ratio / Importance Sampling:

- unbiased estimator of policy gradient

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}_{\rho_{\boldsymbol{\theta}_i}} \left[\frac{\rho_{\boldsymbol{\theta}_k}(\boldsymbol{s},\boldsymbol{a})}{\rho_{\boldsymbol{\theta}_i}(\boldsymbol{s},\boldsymbol{a})} \nabla \log \pi_{\boldsymbol{\theta}_k} A^\pi(\boldsymbol{s}, \boldsymbol{a}) \, (\boldsymbol{a} \, | \boldsymbol{s} \,) \right] \tag{7}$$

- Another off-policy policy gradient estimator simplifies the likelihood ratio term by introducing bias (Degris et al., 2012):

$$\nabla J(\boldsymbol{\theta}) \approx \mathbb{E}_{\rho_{\boldsymbol{\theta}_i}} \left[\frac{\pi_{\boldsymbol{\theta}_k}}{\pi_{\boldsymbol{\theta}_i}} \nabla \log \pi_{\boldsymbol{\theta}_k}(\boldsymbol{a}|\boldsymbol{s}) A^\pi(\boldsymbol{s}, \boldsymbol{a}) \right] \tag{8}$$

Mixture Likelihood Ratio / Multiple Importance Sampling:

- MLR based policy gradient can be obtained by replacing $\frac{\rho_{\boldsymbol{\theta}_k}(\boldsymbol{s},\boldsymbol{a})}{\rho_{\boldsymbol{\theta}_i}(\boldsymbol{s},\boldsymbol{a})}$ or $\frac{\pi_{\boldsymbol{\theta}_k}}{\pi_{\boldsymbol{\theta}_i}}$ with $\frac{\rho_{\boldsymbol{\theta}_k}(\boldsymbol{s}_t,\boldsymbol{a}_t)}{\frac{1}{|U_k|}\sum_{i\in U_k} \rho_{\boldsymbol{\theta}_i}(\boldsymbol{s}_t,\boldsymbol{a}_t)}$ or $= \frac{\pi_{\boldsymbol{\theta}_k}(\boldsymbol{s}_t,\boldsymbol{a}_t)}{\frac{1}{|U_k|}\sum_{i\in U_k} \pi_{\boldsymbol{\theta}_i}(\boldsymbol{s}_t,\boldsymbol{a}_t)}$, where U_k is the reuse set.
- Similar result for **episode**-based approaches: Metelli et al. (2020); Zheng et al. (2021).
- **lower variance** than individual likelihood ratio (LR) estimator and still **unbiased**.

Actor-Critic Method

To estimate the MLR policy gradient, we need to model the value function $V^{\pi}(\boldsymbol{s})$ and policy function $\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})$ using actor-critic method:

- a widely used architecture based on the policy gradient theorem.
- **Actor** corresponds to a action-selection policy $\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})$
- **Critic**: corresponds to a parametric value function $V_{\boldsymbol{w}}(\boldsymbol{s})$

Following (Bhatnagar et al., 2009; Konda and Tsitsiklis, 2003), a typical actor-critic update can be written as

$$\textbf{TD Error}: \quad \delta_k = r_t + \gamma V_{\boldsymbol{w}_k}(\boldsymbol{s}') - V_{\boldsymbol{w}_k}(\boldsymbol{s}) \tag{9}$$

$$\textbf{Critic}: \quad \boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \eta_w \delta_k \nabla_w V_{\boldsymbol{w}_k}(\boldsymbol{s}) \tag{10}$$

$$\textbf{Actor}: \quad \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta_\theta \nabla J(\boldsymbol{\theta}) \tag{11}$$

where η_w and η_θ represent learning rates for critic and actor respectively. The policy gradient $\nabla J(\boldsymbol{\theta})$ is estimated by MLR policy gradient estimate (the previous slide).

Variance Reduced Experience Replay (VRER)

Theorem (Selection Rule)

At the kth iteration where the target distribution is ρ_k, the reuse set U_k includes stationary distributions at ith iteration i.e., ρ_i with $(\boldsymbol{\theta}_i, \boldsymbol{w}_i)$, whose ILR policy gradient estimator's total variance is no greater than c times the total variance of the vanilla PG estimator for some constant $c > 1$. Mathematically,

$$Tr\left(Var\left[\widehat{\nabla\mu}_{i,k}^{ILR}\middle| M_k\right]\right) \leq c\, Tr\left(Var\left[\widehat{\nabla\mu}_{k}^{PG}\middle| M_k\right]\right). \quad (12)$$

Then, based on such reuse set U_k, the total variance of the MLR policy gradient estimator is no greater than $\frac{c}{|U_k|}$ times the total variance of vannila PG estimator,

$$Tr\left(Var\left[\widehat{\nabla\mu}_{k}^{MLR}\middle| M_k\right]\right) \leq \frac{c}{|U_k|} Tr\left(Var\left[\widehat{\nabla\mu}_{k}^{PG}\middle| M_k\right]\right). \quad (13)$$

Remark: $\widehat{\nabla\mu}_k^{PG}$, $\widehat{\nabla\mu}_{i,k}^{ILR}$ and $\widehat{\nabla\mu}_k^{MLR}$ are sample average approximation of vanilla policy gradient, individual likelihood ratio and mixture likelihood ratio based policy gradient respectively.

Algorithm

Input: the selection threshold constant c; the maximum number of iterations K; the number of iterations in offline optimization K_{off}; the number of replications per iteration n_k;
Initialize actor parameter $\boldsymbol{\theta}_1$ and critic parameter $\boldsymbol{w}_1$. Store them in $M_1 = M_0 \cup \{\boldsymbol{\theta}_1, \boldsymbol{w}_1\}$;
for $k = 1, 2, \ldots, K$ **do**

1. Collect transitions $\mathcal{T}_k = \{(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}, r_t)\}_{t=1}^{n_k}$ from real system with $\boldsymbol{\pi}_{\boldsymbol{\theta}_k}$; Update the sets $\mathcal{D}_k \leftarrow \mathcal{D}_{k-1} \cup \mathcal{T}_k$;
2. Initialize $U_k = \emptyset$, screen all historical transitions and associated policies in U_k, and construct the reuse set U_k;

for $(\boldsymbol{\theta}_i, \boldsymbol{w}_i) \in M_k$ *(all models visited utill kth iteration)* **do**

(a) Compute and store the new likelihoods: $\mathcal{L}_k \leftarrow \mathcal{L}_{k-1} \cup \pi_{\boldsymbol{\theta}_k}(\mathcal{D}_k) \cup \pi_{\boldsymbol{\theta}_{[1:k]}}(\mathcal{T}_k)$

(b) Compute $\mathrm{Tr}\left(\mathrm{Var}\left[\left.\widehat{\nabla\mu}_{i,k}^{ILR}\right| M_k\right]\right)$ and $\mathrm{Tr}\left(\mathrm{Var}\left[\left.\widehat{\nabla\mu}_{k}^{PG}\right| M_k\right]\right)$.

if $Tr\left(Var\left[\left.\widehat{\nabla\mu}_{i,k}^{ILR}\right| M_k\right]\right) \leq c\,Tr\left(Var\left[\left.\widehat{\nabla\mu}_{k}^{PG}\right| M_k\right]\right)$ **then**

$U_k \leftarrow U_k \cup \{i\}$.

end

end

3. Reuse the historical samples associated with U_k and stored likelihoods $\mathcal{L}_k$ to update actor and critic:

(a) Let $\boldsymbol{\theta}_k^0 = \boldsymbol{\theta}_k$ and $\boldsymbol{w}_k^0 = \boldsymbol{w}_k$;

for $h = 0, 1, \ldots, K_{off}$ **do**

(b) **TD Error**: $\delta_k^h = r_t + \gamma V_{\boldsymbol{w}_k^h}(\boldsymbol{s}') - V_{\boldsymbol{w}_k^h}(\boldsymbol{s})$;

(c) **Actor Update**: $\boldsymbol{\theta}_k^{h+1} \leftarrow \boldsymbol{\theta}_k^h + \eta_k \widehat{\nabla\mu}_k^{MLR}$;

(d) **Critic Update**: $\boldsymbol{w}_k^{h+1} = \boldsymbol{w}_k^h + \eta_k \delta_k \nabla_w V_{\boldsymbol{w}_k^h}(\boldsymbol{s})$;

end

4. Update the actor and critic: $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k^{K_{off}}$ and $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k^{K_{off}}$;
5. Store them to the set $M_{k+1} = M_k \cup \left\{(\boldsymbol{\theta}_{k+1}, \boldsymbol{w}_{k+1})\right\}$;

end

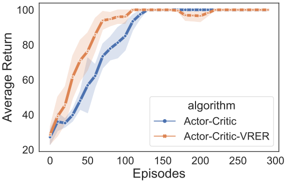
Empirical Study

In the empirical study, we present the experimental evaluation of VRER in combination with **actor critic algorithm** (Bhatnagar et al., 2009) and **proximal policy optimization (PPO)** algorithm (Schulman et al., 2017).

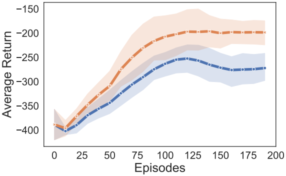
- **Software**: For both actor critic and PPO implementation, we use two open-sourced libraries, Keras and TensorFlow for modeling and automatic differentiation;
- **Control Examples**: (1) *Cartpole* and (2) *Acrobot* control problem from OpenAI gym Brockman et al. (2016).
- **Model structure**: **Actor-Critic** model is composed of a shared initial layer with 128 neurons and separate outputs for the actor and critic. **PPO** algorithm has separate actor and critic neural network models, both of which have two layers with 64 neurons.
  - For the problems with discrete action, we use softmax policy for actor network.
  - For the fermentation problem with a continuous action (feeding rate of substrate), we use the Gaussian policy for actor model.

Github repository: `https://github.com/zhenghuazx/vrer_policy_optimization`

Benchmarks

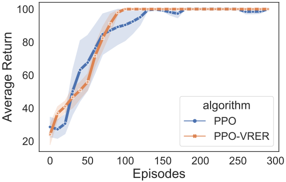


(a) CartPole

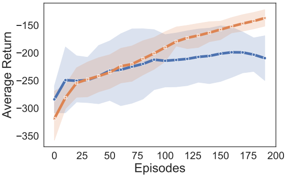


(b) Acrobot

Figure: Convergence results for the Actor-Critic algorithm.



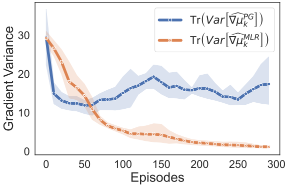
(a) CartPole



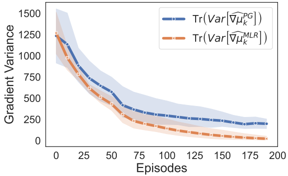
(b) Acrobot

Figure: Convergence results for PPO algorithm.

Result: Lower Variance in Policy Gradient

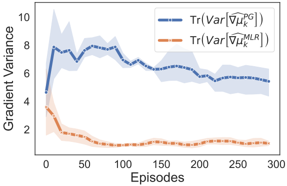


(a) CartPole

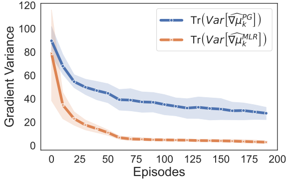


(b) Acrobot

Figure: Convergence results for the Actor-Critic algorithm.



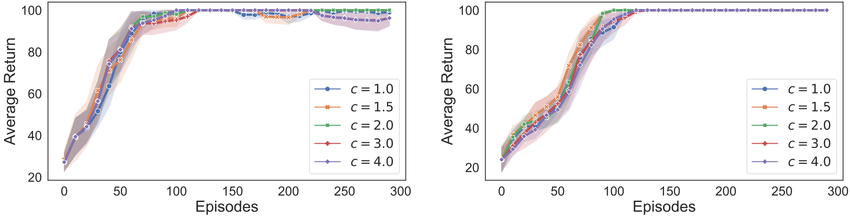
(a) CartPole



(b) Acrobot

Figure: Convergence results for PPO algorithm.

Result: Sensitivity Analysis on c



(a) Actor Critic (b) PPO

Figure: Sensitivity analysis of selection threshold constant c in Cartpole example.

Summary

- Develop VRER to select transitions based on variance reduction;
- Apply mixture likelihood ratio to reduce the variance of off-policy policy gradient;
- Study the applicability of VRER to various actor-critic methods.

Thank you!

Sigrún Andradóttir, Daniel P Heyman, and Teunis J Ott. On the choice of alternative measures in importance sampling with markov chains. *Operations research*, 43(3):509–519, 1995.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. 2016.

Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 179–186, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Mach. Learn.*, 110(9):2419–2468, sep 2021. ISSN 0885-6125. doi: 10.1007/s10994-021-05961-4. URL https://doi.org/10.1007/s10994-021-05961-4.

Vijay R Konda and John N Tsitsiklis. On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.

Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. Importance sampling techniques for policy optimization. *J. Mach. Learn. Res.*, 21:141–1, 2020.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Matthew Schlegel, Wesley Chung, Daniel Graves, Jian Qian, and Martha White. Importance resampling for off-policy prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8 (3):229–256, 1992.

Hua Zheng, Wei Xie, and M Ben Feng. Green simulation assisted policy gradient to accelerate stochastic process control. *arXiv preprint arXiv:2110.08902*, 2021.