Reinforcement Learning with General Utilities: Simpler Variance Reduction and Large State-Action Space

Anas Barakat

Joint work with Ilyas Fatkhullin and Niao He

ICML 2023

ETH zürich

RL with General Utilities - Motivation

- **Constrained RL, risk-sensitive/averse RL:** Conditional Value-at-Risk.
- **Imitation Learning:** f-divergence minimization between state-action occupancy measures of an agent and an expert.
- **Pure exploration:** State visitation entropy maximization.
- **Active Exploration:** Experiment design in Markov Chains.
- ...

Problem Formulation

- MDP $M(\mathcal{S}, \mathcal{A}, \mathcal{P}, F, \rho, \gamma)$ with a general utility function F,
- Parametrized policy $\pi_\theta, \theta \in \mathbb{R}^d$,
- State-action occupancy measure λ^{π_θ}:

$$\lambda^{\pi_\theta}(s, a) \stackrel{\text{def}}{=} \sum_{t=0}^{+\infty} \gamma^t \mathbb{P}_{\rho,\pi_\theta}(s_t = s, a_t = a) .$$

Problem

$$\max_{\theta \in \mathbb{R}^d} F(\lambda^{\pi_\theta})$$

Recent Related Work

- Convex RL and unified problem formulation [Hazan et al., 2019, Zahavy et al., 2021, Zhang et al., 2020, Zhang et al., 2021]
- Direct policy search method [Zhang et al., 2021], Bellman equations being invalidated due to nonlinearity.

Solving RL with General Utilities using Standard RL

$$\nabla_\theta F(\lambda^{\pi_\theta}) = \nabla_\theta V^{\pi_\theta}(r)|_{r=\nabla_\lambda F(\lambda^{\pi_\theta})},$$

where $V^{\pi_\theta}(r) \overset{\text{def}}{=} \mathbb{E}_{\rho,\pi_\theta}\left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t)\right]$.

- double-loop variance-reduced PG method with gradient truncation to control IS weights in the tabular setting.

1. Simpler Variance Reduction

Limitations in Prior Work and our Contributions

1.
   - **Prior work:** double-loop PG algorithm with large batches for variance reduced stochastic PG and parameter knowledge.
   - **Contribution:** single-loop normalized PG using a single trajectory per iteration and reducing parameter knowledge requirements, inspired by [Cutkosky and Orabona, 2019].

2.
   - **Prior work:** Most prior work makes the unrealistic assumption of bounded IS weights variance (in standard RL).
   - **Contribution:** Normalized gradient update guarantees bounded IS weights for softmax (and Gaussian) policies.

1. Simpler Variance Reduction

Algorithm 1 N-VR-PG(General Utilities)

Input: θ_0, T, H, $\{\eta_t\}_{t\geq 0}$, $\{\alpha_t\}_{t\geq 0}$.
Sample τ_0 of length H from M and π_{θ_0}
$\lambda_0 = \lambda(\tau_0, \theta_0)$; $r_0 = \nabla_\lambda F(\lambda_0)$; $r_{-1} = r_0$
$d_0 = g(\tau_0, \theta_0, r_0)$
$\theta_1 = \theta_0 + \alpha_0 \frac{d_0}{\|d_0\|}$
for $t = 1, \ldots, T-1$ **do**
  Sample τ_t of length H from MDP $\mathbb{M}$ and π_{θ_t}
  $u_t = \lambda(\tau_t)(1 - w(\tau_t|\theta_{t-1}, \theta_t))$
  $\lambda_t = \eta_t \lambda(\tau_t) + (1-\eta_t)(\lambda_{t-1} + u_t)$
  $r_t = \nabla_\lambda F(\lambda_t)$
  $v_t = g(\tau_t, \theta_t, r_{t-1}) - w(\tau_t|\theta_{t-1}, \theta_t) g(\tau_t, \theta_{t-1}, r_{t-2})$
  $d_t = \eta_t g(\tau_t, \theta_t, r_{t-1}) + (1-\eta_t)(d_{t-1} + v_t)$
  $\theta_{t+1} = \theta_t + \alpha_t \frac{d_t}{\|d_t\|}$
end for

1. Simpler Variance Reduction

Theoretical Guarantees

- **Challenge:** coupled recursive estimation errors for stochastic PG and occupancy measure VR estimates.

Theorem (Sample complexities)

Under regularity assumptions on the softmax parametrization and the utility function F,

- $\tilde{\mathcal{O}}(\varepsilon^{-3})$ samples to reach an ε-stationary point of the objective function ,
- If F is further concave and the policy overparametrized, $\tilde{\mathcal{O}}(\varepsilon^{-2})$ samples to reach an ε-globally optimal policy .

Simulations

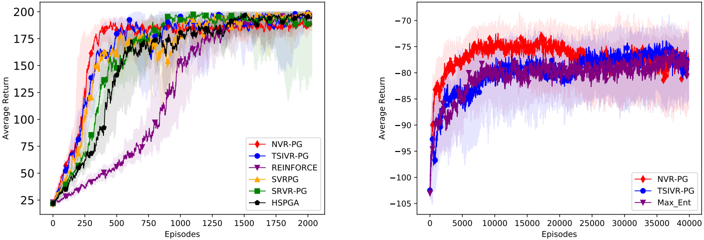

Figure 1: (right) Nonlinear objective maximization in the FrozenLake environment; (left) Standard RL in the CartPole environment.

2. Large State-Action Space Setting

Limitations and Contributions

- **Prior work:** Tabular setting for state-action occupancy measure estimation.
- **Contribution:**

Linear function approximation of the occupancy measure

$$\lambda^{\pi_\theta}(s,a) \approx \langle \phi(s,a), \omega_\theta \rangle\,, \quad \omega_\theta \in \mathbb{R}^m\,, m << |\mathcal{S}| \times |\mathcal{A}|\,.$$

Linear regression procedure

- K steps of SGD over the following objective:

$$L_\theta(\omega) \stackrel{\text{def}}{=} \mathbb{E}_{s\sim\rho, a\sim\mathcal{U}(\mathcal{A})}[(\lambda^{\pi_\theta}(s,a) - \langle \phi(s,a), \omega \rangle)^2]\,,$$

- using Monte-Carlo estimates for the targets $\lambda^{\pi_\theta}(s,a)$ for each state-action pair sampled at each step $k \leq K$.

2. Large State-Action Space Setting

Theoretical Guarantees

Theorem (Sample complexity)

Under

1. regularity of the utility function F,
2. smoothness of the policy parametrization,
3. standard assumptions on the feature map,

stochastic PG with the linear regression subroutine requires

$$\tilde{\mathcal{O}}(\varepsilon^{-4}) \quad \text{samples}$$

to guarantee an ε-first-order stationary point of the objective function up to a function approximation error floor.

Thank you for your attention

Contact: **barakat9anas@gmail.com**

References I

Cutkosky, A. and Orabona, F. (2019).
Momentum-based variance reduction in non-convex sgd.
Advances in neural information processing systems, 32.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. (2019).
Provably efficient maximum entropy exploration.
In *International Conference on Machine Learning*, pages 2681–2691. PMLR.

Zahavy, T., O'Donoghue, B., Desjardins, G., and Singh, S. (2021).
Reward is enough for convex mdps.
Advances in Neural Information Processing Systems, 34:25746–25759.

Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. (2020).
Variational policy gradient method for reinforcement learning with general utilities.
Advances in Neural Information Processing Systems, 33:4572–4583.

Zhang, J., Ni, C., Szepesvari, C., Wang, M., et al. (2021).
On the convergence and sample efficiency of variance-reduced policy gradient method.
Advances in Neural Information Processing Systems, 34:2228–2240.