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

ETH zürich

ICML International Conference On Machine Learning

Anas Barakat Ilyas Fatkhullin Niao He

Motivation

RL with general utilities

- Imitation Learning
- Pure exploration
- Risk-sensitive/averse RL
- Active exploration for experimental design
- ...

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:

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

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

[1, 2, 3, 4]

Policy gradient theorem [4]

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

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

Challenges

- double-loop, large batch, params
- occupancy measure estimation in **large** state-action space

Normalized Variance Reduced PG for RL with General Utilities

Algorithm 1 N-VR-PG (General Utilities)

Input: $\theta_0, T, H, \{\eta_t\}_{t\geq 0}, \{\alpha_t\}_{t\geq 0}$.
for $t = 1, \ldots, T-1$ **do**
  Sample τ_t of length H from MDP 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) single-loop batch free; (2) normalization implies boundedness of IS weights

Sample complexity for N-VR-PG

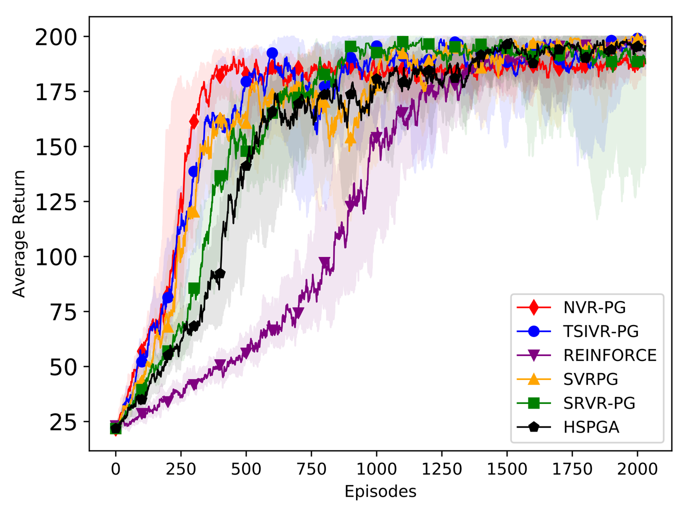
Under smoothness conditions on F and softmax π_θ,

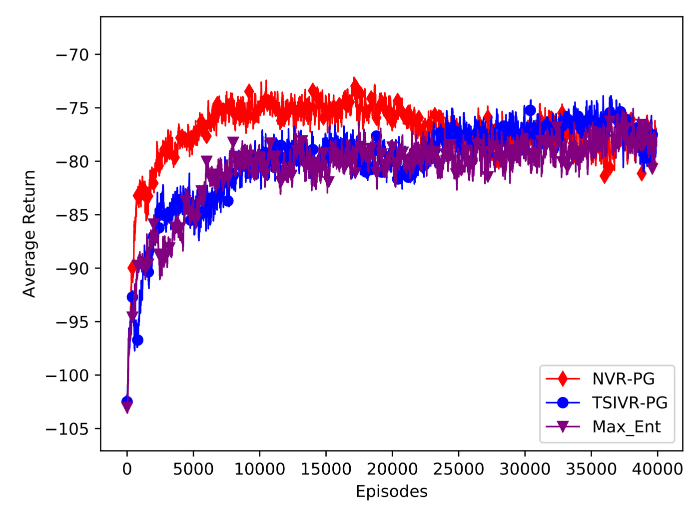
Setting	Guarantee	Sample complexity
F non-concave	$\mathbb{E}[\|\nabla F(\lambda^{\pi_{\theta_{\text{out}}}})\|] \leq \varepsilon$	$\tilde{\mathcal{O}}(\varepsilon^{-3})$
F concave*	$\mathbb{E}[F^* - F(\lambda^{\pi_{\theta_{\text{out}}}})] \leq \varepsilon$	$\tilde{\mathcal{O}}(\varepsilon^{-2})$

$*$ with overparametrized softmax policy

Simulations

(left) standard RL in CartPole; (right) nonlinear obj. maximization in FrozenLake





Large state-action space

- Linear function approximation of the occupancy measure

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

- Linear regression procedure:
  - K steps of SGD over the objective:

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

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

Stochastic PG with Linear Occupancy Measure Approximation

Algorithm 2 Stochastic PG with Linear Function Approximation

Input: $\theta_0 \in \mathbb{R}^d, T, N \geq 1, \alpha > 0, K \geq 1, \beta > 0, H$.
for $t = 0, \ldots, T-1$ **do**
  Run SGD for K steps of linear regression to obtain $\hat{\omega}_{\theta_t}$.
  Define $r_t = \nabla_\lambda F(\hat{\lambda}_t)$ where $\hat{\lambda}_t(\cdot,\cdot) = \langle \phi(\cdot,\cdot), \hat{\omega}_{\theta_t} \rangle$.
  Sample N independent trajectories $(\tau_t^{(i)})_{1\leq i\leq N}$ of length H with π_{θ_t}
  $\theta_{t+1} = \theta_t + \frac{\alpha}{N} \sum_{i=1}^N g(\tau_t^{(i)}, \theta_t, r_{t-1})$
end for
Return: θ_T

Sample complexity

Assumptions: (a) regularity of the utility function F, (b) smoothness of π_θ and (c) standard assumptions on the feature map ϕ.

Theorem: Stochastic PG with 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**.

References

[1] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. *ICML 2019*.
[2] Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. *NeurIPS 2021*.
[3] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *NeurIPS 2020*.
[4] Junyu Zhang, Chengzhuo Ni, Zheng Yu, Csaba Szepesvari, and Mengdi Wang. On the convergence and sample efficiency of variance-reduced policy gradient method. *NeurIPS 2021*.