On the Global Optimality of Policy Gradient Methods in General Utility Reinforcement Learning

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Anas Barakat **Souradip Chakraborty** **Peihong Yu** **Pratap Tokekar** **Amrit Singh Bedi**

UNIVERSITY OF MARYLAND 18 56 — UNIVERSITY OF CENTRAL FLORIDA 1963

General Utility RL

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

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

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

Applications

Imitation Learning Pure exploration

Risk-sensitive/averse RL Experiment design

...

Gaps in Prior Work on Theoretical RLGU

- Different analysis from recent PG advances.
- Convex RL/RLGU, mainly restricted to **TABULAR** setting.

Global Optimality of PG in RLGU?

1. How to reconcile PG analysis in standard RL with RLGU?
2. How to scale RLGU policy optimization to **large scale** $\mathcal{S} \times \mathcal{A}$?
   - How to estimate unknown λ^{π_θ} in **large scale** $\mathcal{S} \times \mathcal{A}$?

Challenges

1. Nonconvex optimization problem (but hidden convexity).
2. No Forward Bellman equation, rather Backward Bellman flow.
3. Monte-Carlo (count-based) estimates do not scale.

Main Contributions

1. **Tabular setting:** RLGU objective gradient domination.
2. **Function Approx.setting:** PG algorithm and sample complexity scaling with dimension of the function approx. params.

Occupancy Measure Estimation via Function Approx.

- Motivation: $d^\pi(s)$ linear in density features in low-rank MDPs:

$$P(s'|s,a) = \langle \phi(s,a), \mu(s') \rangle \implies d^\pi(s) = \rho(s) + \langle \omega_\pi, \mu(s) \rangle\,,$$

- Samples $\{s_i\}_{i=1}^n \in \mathcal{S}^n$ and function approximation class:

$$\Lambda := \{ p_\omega \in \Delta(\mathcal{S}) \,|\, \omega \in \Omega \subseteq \mathbb{R}^m \}\,, \quad m \ll |\mathcal{S}|\,.$$

MLE : $\quad d^{\pi_\theta} \simeq p_{\omega^*}\,, \qquad \omega^* \in \arg\max_{\omega \in \Omega} \frac{1}{n} \sum_{i=1}^n \log p_\omega(s_i)\,.$

Algorithm (PG for RLGU with Occupancy Approx)

for $t = 0, \ldots, T-1$ **do**

1/Pseudo-reward learning via Occup. approx.

Compute MLE estimator $\hat{\lambda}_t = \hat{d}^{\pi_{\theta_t}} \cdot \pi_{\theta_t}$, then define:

$$\hat{r}_t = \nabla_\lambda F(\hat{\lambda}_t) \simeq \nabla_\lambda F(\lambda^{\pi_{\theta_t}})$$

2/Policy parameter update

Sample N independent H-trajectories $(\tau_t^{(i)})_{1 \le i \le N}$ using π_{θ_t}.

$$\theta_{t+1} = \theta_t + \alpha \widehat{\nabla_\theta F(\lambda^{\pi_{\theta_t}})}, \quad \widehat{\nabla_\theta F(\lambda^{\pi_{\theta_t}})} = \frac{1}{N} \sum_{i=1}^N g(\tau_t^{(i)}, \theta_t, \hat{r}_t).$$

Return: θ_T

RLGU Gradient Domination (Tabular Setting)

Theorem. Under concavity of F w.r.t. λ, for every $\theta \in \mathbb{R}^d$,

$$\underbrace{F(\lambda^{\pi_{\theta^\star}}) - F(\lambda^{\pi_\theta})}_{\text{Optimality Gap}} \le \kappa(\theta) \underbrace{\max_{\bar{\pi} \in \Pi} \langle \bar{\pi} - \pi_\theta, \nabla_\theta F(\lambda^{\pi_\theta}) \rangle}_{\text{First-Order Stationarity Metric}}\,,$$

with policy-dependent mismatch coefficient:

$$\kappa(\theta) := \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^\star(r_\theta)}}{\mu} \right\|_\infty\,, \quad r_\theta := \nabla_\lambda F(\lambda^{\pi_\theta}) \quad d = |\mathcal{S}| \cdot |\mathcal{A}|\,,$$

$\pi^\star(r) \in \arg\max_{\pi \in \Pi} V^\pi(r)$ for any $r \in \mathbb{R}^{|\mathcal{S}| \cdot |\mathcal{A}|}$, $\theta^\star$ an optimal policy parameter and any $\mu \in \Delta(\mathcal{S})$ s.t. $\mu(s) > 0$ for all states $s \in \mathcal{S}$.

Assumptions

1. Function approx. class regularity (parameter compactness $B_\omega := \max_{\omega \in \Omega} \|\omega\|_\infty$, realizability $\lambda^\pi \in \Lambda$, Lipschitzness of p_ω)
2. Smooth policy (over)parametrization (softmax π_θ).
3. Utility smoothness ($\|\nabla_\lambda F(\lambda_1) - \nabla_\lambda F(\lambda_2)\|_2 \le L_\lambda \|\lambda_1 - \lambda_2\|_2$.)

Sample complexity (Function Approximation Setting)

MLE: $\forall \delta$, w.p. at least $1-\delta$, $\|\hat{\lambda}^{\pi_\theta} - \lambda^{\pi_\theta}\|_1 \le 6\sqrt{\frac{12\, m \log\left(\frac{2[B_\omega B_L n]}{\delta}\right)}{n}}$.

Setting	Guarantee	Sample complexity
F non-concave	$\mathbb{E}[\|\nabla F(\lambda^{\pi_{\theta_{\text{out}}}})\|^2] \le \varepsilon$	$\tilde{\mathcal{O}}(\underbrace{m}_{\text{dim(fun. approx)}} \varepsilon^{-2})$
F **concave**	$\mathbb{E}[F^* - F(\lambda^{\pi_{\theta_{\text{out}}}})] \le \varepsilon$	$\tilde{\mathcal{O}}(\underbrace{m}_{\text{dim(fun. approx)}} \varepsilon^{-4})$

Future Work

- Continuous state-action spaces
- Beyond discounted infinite-horizon rewards
- Convex Markov Games

PAPER

References

[1] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. JMLR, 2021.
[2] J. Zhang, C. Ni, C. Szepesvari, and M. Wang. On the convergence and sample efficiency of variance-reduced policy gradient method. In NeuRIPS, 2021.
[3] A. Barakat, I. Fatkhullin, and N. He. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In ICML, 2023.