Policy Gradient: Optimal Estimation, Convergence, and Generalization beyond Cumulative Rewards

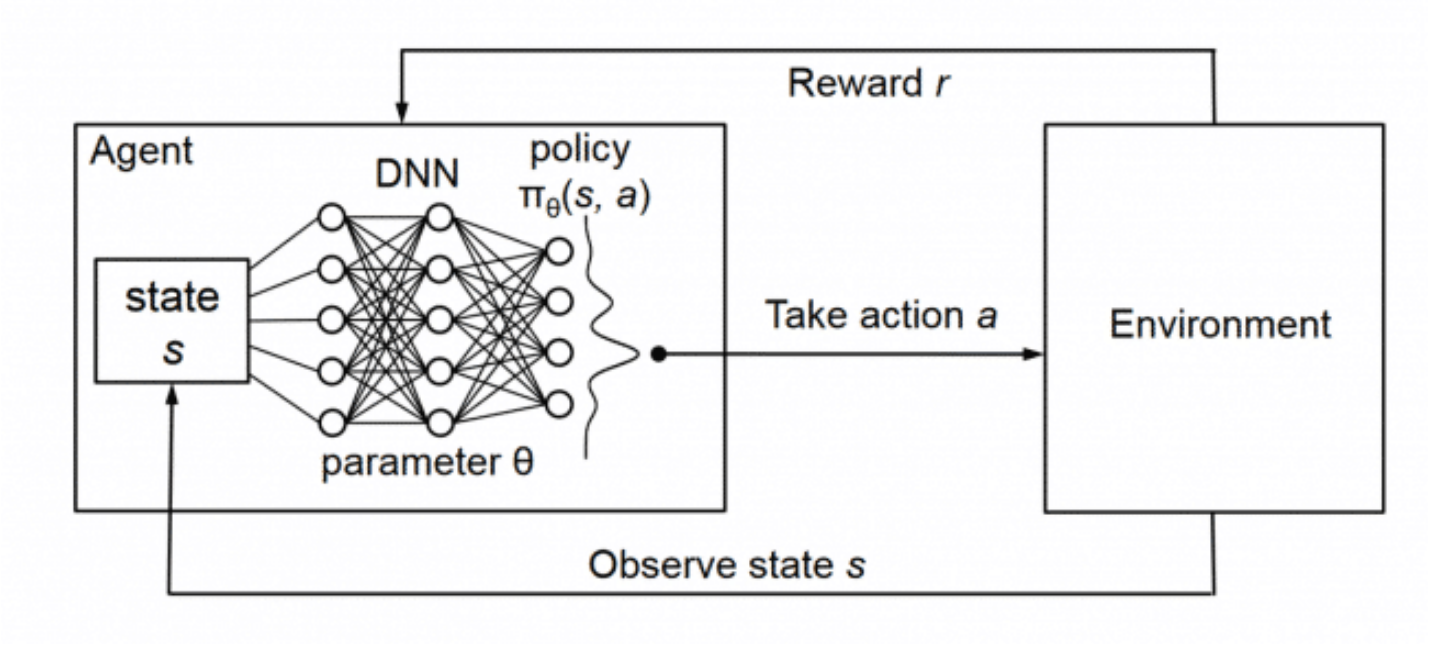
Joint work with Chengzhuo Ni (Princeton), Xuezhou Zhang (Princeton), Ruiqi Zhang (PKU), Junyu Zhang (NUS), Alec Koppel (Amazon), Amrit Singh Bedi (ARO), Csaba Szepesvari (DeepMind)

Mengdi Wang

February, 2022

Policy-based Method

Direct policy optimization: search for the best θ^* via gradient ascent



Policy gradient theorem (Sutton, 2000):

$$\nabla_\theta V^{\pi_\theta} = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t, a_t) \cdot \nabla_\theta \log \pi_\theta(a_t|s_t)\right].$$

Policy Gradient: Questions To Ask

Estimation:

- Off-policy distribution shift brings large bias & variance
  The most commonly used method is still importance sampling: exponentially large variance
- Leverage function approximation for more accurate PG?
  Tons of value-based methods for learning $\hat{Q}$. Can we learn $\nabla_\theta Q$ in a similar way?
- Minimax optimal estimation for (off-policy) PG estimation?

Use PG for policy optimization:

- Nonconvexity and global convergence *(Agarwal et al 19, Mei et al 20)*
- Sample efficiency for finding optimal policy: use a better PG estimator? *(a long list of variance-reduced methods)*
- More generally: can PG method still work beyond standard RL?

Policy Gradient Bellman Equation (Ni et al. 22)

Bellman equation gives:

$$Q_h^\theta = r_h + \mathcal{P}_{\theta,h} Q_{h+1}^\theta, \quad h \in [H]$$

Differentiating both sides of the Bellman equation, we get

$$\nabla_\theta Q_h^\theta = \mathcal{P}_{\theta,h} \left(\left(\nabla_\theta \log \Pi_{\theta,h+1} \right) Q_{h+1}^\theta + \nabla_\theta Q_{h+1}^\theta \right), \quad h \in [H]$$

where we define the operator $\nabla_\theta \log \Pi_{\theta,h}$ by

$$\left(\left(\nabla_\theta \log \Pi_{\theta,h} \right) f \right)(s,a) := \left(\nabla_\theta \log \pi_{\theta,h}(a|s) \right) f(s,a).$$

Hint: One can use function approximation to estimate Q and $\nabla_\theta Q$ in a similar way

Double Fitted Iteration for Policy Gradient Estimation

Given a function class $\mathcal{F}$ and batch data $\{(s_h^k, a_h^k)\}$ of K i.i.d. episodes, we apply iterative function fitting (in the same spirit with Fitted Q Iteration/Evaluation):

$$\widehat{Q}_h^{\theta,\mathrm{FPG}} = \arg\min_{f\in\mathcal{F}} \left[\lambda\rho(f) + \sum_{k=1}^{K}\left(f\left(s_h^{(k)}, a_h^{(k)}\right) - r_h^{(k)} - \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a' \middle| s_{h+1}^{(k)}\right) \widehat{Q}_{h+1}^{\theta,\mathrm{FPG}}\left(s_{h+1}^{(k)}, a'\right) \mathrm{d}a'\right)^2\right] \quad (2)$$

$$\widehat{\nabla_\theta^j Q_h^{\theta,\mathrm{FPG}}} = \arg\min_{f\in\mathcal{F}} \left[\lambda\rho(f) + \sum_{k=1}^{K}\left(f\left(s_h^{(k)}, a_h^{(k)}\right) - \int_{\mathcal{A}} \pi_{\theta,h+1}\left(a' \middle| s_{h+1}^{(k)}\right)\left(\left(\nabla_\theta^j \log \pi_{\theta,h+1}\left(a' \middle| s_{h+1}^{(k)}\right)\right) \widehat{Q}_{h+1}^{\theta,\mathrm{FPG}}\left(s_{h+1}^{(k)}, a'\right) + \widehat{\nabla_\theta^j Q_{h+1}^{\theta,\mathrm{FPG}}}\left(s_{h+1}^{(k)}, a'\right)\right)\mathrm{d}a'\right)^2\right] \quad (3)$$

Theorem

Double Fitted PG Iteration = Plug-In Model-Based Estimator

Linear MDP: $\widehat{\nabla_\theta Q^\theta} = \nabla_\theta \widehat{Q^\theta}$

Optimal Statistical Error Bounds

Given off-policy data:

Theorem ((Variance-Aware) Finite-Sample Error Bound)

With high-probability,

$$|\langle t, \widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta \rangle| \leq \sqrt{\frac{2t^\top \Lambda_\theta t}{K} \cdot \log \frac{8}{\delta}} + O(\frac{1}{K})$$

where Λ_θ is the error covariance (formula too long to include)

Theorem

Cramer-Rao Lower Bound Any unbiased estimator have variance at least $\frac{1}{K}\Lambda_\theta$

Distribution shift and function approximation are coupled together

Off-policy PG estimation: suppose the batch data and target policy have state-action occupancy measure $\bar{\mu}$, μ^θ

Theorem (Finite-Sample Error Bound by Distribution Shift)

With high-probability: $|\widehat{\nabla_\theta v_\theta} - \nabla_\theta v_\theta|_\infty \leq 4H^{2.5}\sqrt{\frac{1+\chi_{\mathcal{F}}(\mu^\theta,\bar{\mu})}{K}} + \tilde{O}\left(\frac{1}{K}\right).$

$\mathcal{F}$-restricted chi-square: Measuring the distribution shift in the function class

$$\chi^2_{\mathcal{F}}(p_1, p_2) := \sup_{f\in\mathcal{F}} \frac{\mathbb{E}_{p_1}[f(x)]^2}{\mathbb{E}_{p_2}[f(x)^2]} - 1$$

- $\chi^2_{\mathcal{F}}(\mu^\theta, \bar{\mu}) \ll \chi^2(\mu^\theta, \bar{\mu}) \ll \|\mu^\theta/\bar{\mu}\|_\infty$
- In linear MDP: $\chi^2_{\mathcal{F}} \leq$ relative condition number
- $O(\sqrt{\chi^2_{\mathcal{F}}(\mu^\theta, \bar{\mu})/K})$ is minimax optimal

Fundamental Property of RL: Rewards are Cumulative

Rewards are additive, therefore:

- Every state can be assessed via a value function:

$$V(s) = \mathbb{E}[r(s_1) + \gamma \cdot r(s_2) + \cdots | s_1 = s]$$

- Bellman equation (Bellman, 1945) holds:

$$V(s) = r(s) + \gamma \mathbb{E}[V(s') | s]]$$

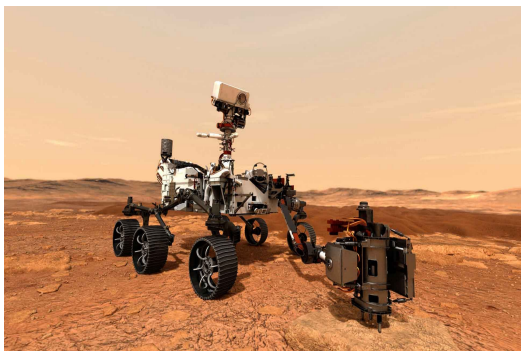
- Policy gradient theorem (Sutton, 2000) holds:

$$\nabla_\theta V^{\pi_\theta} = \mathbb{E}^{\pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t, a_t) \cdot \nabla_\theta \log \pi_\theta(a_t | s_t) \right].$$

The cumulative nature of rewards is key to all RL algorithms

RL with general utilities

- Consider Markov Decision Process: $\mathrm{MDP}(\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$.
- Problems beyond cumulative reward?



(a) Exploration

(b) Risk aversion

(c) Imitation

RL with general utilities

- Maximizing a policy's long term utility:

$$\underset{\theta}{\text{maximize}} \quad R(\pi_\theta) := F(\lambda^{\pi_\theta})$$

  - π_θ the policy, parameterized by θ.
  - λ^π the unnormalized state-action occupancy measure.

$$\lambda^\pi_{sa} := \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{P}\Big(s_t = s, a_t = a \,\Big|\, \pi, s_0 \sim \xi\Big).$$

  - F a concave function.

- For concave F, it is sufficient to explore over stationary policies.

General Utilities for RL

- cumulative reward, linear F:

$$F(\lambda^{\pi_\theta}) = \langle \text{occupancy measure}, \text{reward} \rangle.$$

- exploration over state space:

$$F(\lambda^{\pi_\theta}) = \text{Entropy}\big(\text{state visitation frequency}\big)$$

- exploration over the feature space:

$$F(\lambda^{\pi_\theta}) = \sigma_{\min}\big(\text{covariance matrix}\big).$$

- Imitation:

$$F(\lambda^{\pi_\theta}) = -D_{KL}\big(\text{occupancy measure} \,\big|\big|\, \text{some distribution}\big)$$

Moving beyond cumulative rewards is hard

- Difficulty: the Bellman equation, value function, q function, dynamic programming, all fail.
- Questions:
  - Is policy search still viable?
  - If so, can we do policy search in parameter space? to handle large state-action space.
- This is important for deriving scalable parameterized algorithms for large scale RL problems.

What are the existing results?

- RL utilities beyond cumulative rewards: Max entropy exploration (Hazan et al., 2019); Imitation (Schaa, 1997), (Argall et al., 2008)...; Constrained RL: (Eitan Altman, 1999), (Achiam et al., 2017) ...
  - Many of them does not allow function approximation.
  - We provide a general solution to these problems.

- Policy gradient: (Sutton et al., 2000), (Pirotta et al., 2015)...
  - limited to cumulative rewards
  - convergence to stationary point

- Recently efforts on PG method for cumulative rewards, convergence to global optima: (Agarwal et al., 2019), (Mei et al., 2020)...
  - We guarantee global optimality for more general utilities, via novel perspective of hidden convexity.

What's the policy gradient for general utilities?

- Policy gradient theorem (Sutton et al., 2000), cumulative reward:

$$\nabla_\theta V^{\pi_\theta} = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t, a_t) \cdot \nabla_\theta \log \pi_\theta(a_t|s_t)\right].$$

  It fails for general utilities since Q-function isn't well-defined.

- For general utilities, by chain rule

$$\nabla_\theta R(\pi_\theta) = \sum_{s,a} \frac{\partial F(\lambda^{\pi_\theta})}{\partial \lambda_{sa}} \cdot \nabla_\theta \lambda_{sa}^{\pi_\theta}.$$

- Both $\frac{\partial F(\lambda^{\pi_\theta})}{\partial \lambda_{sa}}$ and $\nabla_\theta \lambda_{sa}^{\pi_\theta}$ are hard to estimate.

What's the policy gradient for general utilities?

Theorem (Variational Policy Gradient Theorem)

$$\nabla_\theta R(\pi_\theta) = \lim_{\delta \to 0_+} \operatorname*{argmax}_x \inf_z \left\{ V(\theta; z) + \delta \nabla_\theta V(\theta; z)^\top x - F^*(z) - \frac{\delta}{2} \|x\|^2 \right\}.$$

- F^*: convex conjugate of F.
- z: the shadow reward.
- $V(\theta; z)$: cumulative reward with reward function z, policy π_θ.

Estimation of Variational PG

On-policy estimation via stochastic min-max optimization:

1. On policy sampling: Generate episodic sample paths ζ_i, $i \in [n]$ using π_θ, where $\zeta_i = \{(s_t^{(i)}, a_t^{(i)})\}$
2. For any z, estimate $V(\theta, z)$ and $\nabla V(\theta, z)$ by:

$$\tilde{V}(\theta; z) := \frac{1}{n}\sum_{i=1}^{n} V(\theta; z; \zeta_i) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \gamma^k \cdot z(s_k^{(i)}, a_k^{(i)}),$$

$$\nabla\tilde{V}(\theta; z) := \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta V(\theta; z; \zeta_i) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{a\in\mathcal{A}} \gamma^k \cdot Q(s_k^{(i)}, a; z)\nabla_\theta \pi_\theta(a|s_k^{(i)}).$$

3. Plug-in estimation (with $\delta \approx 0$):

$$\hat{\nabla}_\theta R(\pi_\theta; \delta) := \operatorname*{argmax}_x \inf_{\|z\|_\infty \le \ell_F} \left\{ -F^*(z) + \tilde{V}(\theta; z) + \delta\nabla_\theta \tilde{V}(\theta; z)^\top x - \frac{\delta}{2}\|x\|^2 \right\}$$

Estimation of Variational PG: Statistical Bound

Theorem

Under smoothness assumptions, we have

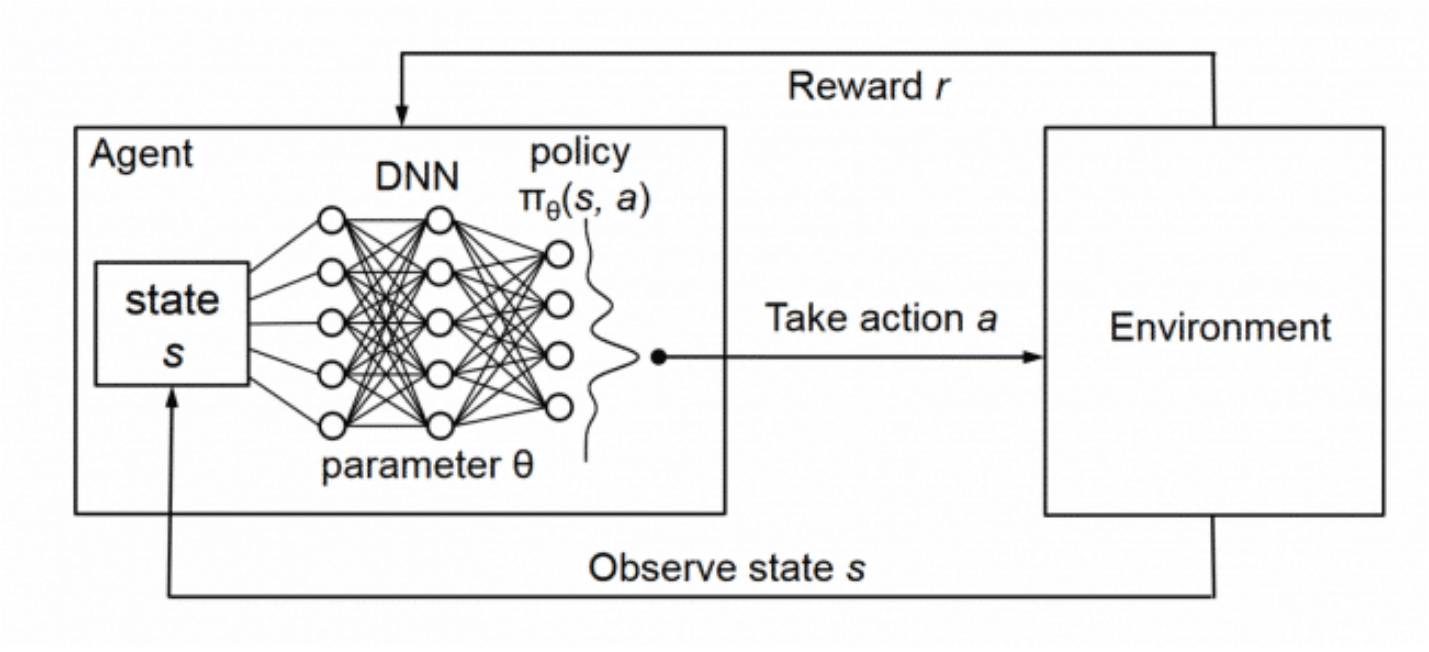
$$\mathbb{E}[\|\hat{\nabla}_\theta R(\pi_\theta) - \nabla_\theta R(\pi_\theta)\|^2] \leq \frac{C}{(1-\gamma)^6 n}.$$

Proof idea: analyze the stochastic stability of saddle points

Yes, we can estimate PG for general utilities with polynomial sample complexity.

Landscape of the nonconvex utility

Recall:



- $\max_\theta R(\pi_\theta)$ is highly nonconvex: saddle points, bad local optimas.
- Nonconvex even for standard RL

Landscape of the nonconvex utility

Hidden Convexity: However, the problem $\max_\pi R(\pi)$ is equivalent to

$$\max_\lambda F(\lambda) \text{ subject to } \lambda \in \text{Polyhedron}$$

Theorem

Under proper assumptions (bijection/overparametrization), every first-order stationary solution of the (possibly nonsmooth) nonconvex problem

$$\max_\theta R(\pi_\theta)$$

is a global optimal solution*.*

Rate of convergence to global optima

Theorem

Consider the policy gradient update

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta R(\pi_{\theta_t}).$$

Under proper assumptions, the policy gradient update satisfies

$$R(\pi_{\theta^*}) - R(\pi_{\theta_t}) \leq \mathcal{O}\big(1/t\big).$$

Additionally, if $F(\cdot)$ is strongly concave, we have

$$R(\pi_{\theta^*}) - R(\pi_{\theta_t}) \leq \mathcal{O}\big(\exp\{-\alpha \cdot t\}\big), \quad \alpha \in (0, 1).$$

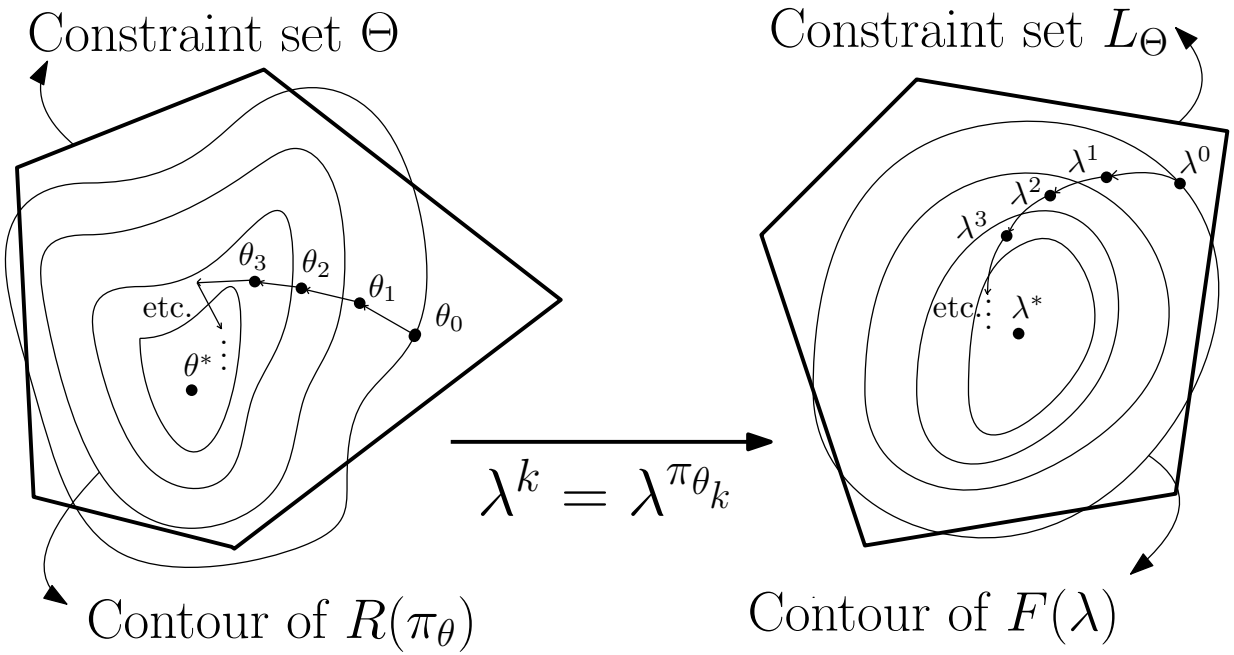
- For tabular MDP, no parameterization: $\mathcal{O}(1/\epsilon)$ iteration complexity.
- Improving the $\mathcal{O}(1/\epsilon^2)$ state-of-the-art result.

Rate of convergence to global optima

- Key intuition behind: **hidden convexity**:

$$\max_{\theta \in \Theta} R(\pi_\theta) \quad \Longleftrightarrow \quad \max_{\lambda \in \mathcal{L}} F(\lambda).$$

- Gradient flow in θ space $\Longleftrightarrow$ "gradient flow" in λ space.



Consequence for cumulative reward

- Cumulative reward, no parameterization (tabular MDP)

$$\max_{\pi} V^{\pi} \quad \Longleftrightarrow \quad \max_{\lambda} \langle r, \lambda \rangle \ s.t. \ \lambda \in \mathcal{L}.$$

  where $\mathcal{L}$ is a **polyhedron**.

- All assumptions can be verified.
- Iteration complexity $\mathcal{O}\left(\frac{1}{\epsilon}\right)$, improve the state-of-the-art.

Sample Complexity for Policy Optimization

Theorem (Putting estimation and convergence together (Informal))

There exists an algorithm for $\max_\theta R(\pi_\theta)$ *that finds an* ϵ*-optimal policy using* $O(1/\epsilon^2)$ *sample trajectories.*

Remarks:

- Optimal dependence on ϵ
- Requires smoothness conditions
- In short: Sample complexity for General-Utility RL is comparable to stochastic convex optimization

Summary

PG estimation via Double Fitted Iteration

- Policy gradient Bellman equation
- Minimax optimal PG estimation of from on/off-policy data
- Statistical error determined by the coupling of distribution shift and function approximation

Beyond cumulative rewards

- Variational PG Theorem: works when Bellman equation fails
- PG estimation via stochastic minimax optimization.

Convergence to global opt (cumulative reward and beyond)

- Exploit the hidden convexity in the occupancy measure.
- $1/\epsilon$ iteration complexity and $1/\epsilon^2$ sample complexity