Variational Policy Gradient Method for Reinforcement Learning with General Utilities

Junyu Zhang, Alec Koppel, Amrit Singh Bedi,

Csaba Szepesvari, Mengdi Wang

Conference on Neural Information Processing Systems (NeurIPS)

Oct 18, 2020

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

- More examples...

RL with general utilities

- Maximizing a policy's long term utility:

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

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.

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:

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

General Utilities for RL

- cumulative reward, linear F:

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

- exploration over state space:

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

General Utilities for RL

- cumulative reward, linear F:

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

- exploration over state space:

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

- exploration over the feature space:

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

General Utilities for RL

- cumulative reward, linear F:

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

- exploration over state space:

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

- exploration over the feature space:

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

- 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.

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.

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.

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

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.

Whats 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.

Whats 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.

Whats the policy gradient for general utilities?

Theorem (Variational Policy Gradient Theorem)

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

Whats 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.

Landscape of the nonconvex utility

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

Theorem

Under proper assumptions, 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

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

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

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

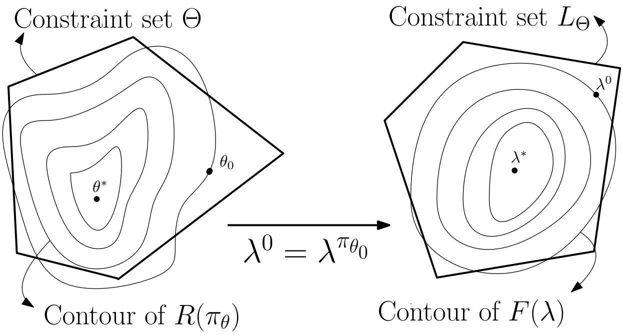
- For tabular MDP, no parameterization: $\mathcal{O}(1/\epsilon)$ iteration complexity.
- Improving the $\mathcal{O}(1/\epsilon^2)$ state-of-the-art result.

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.

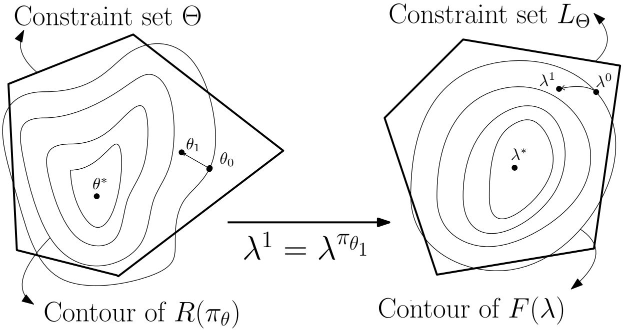


Rate of convergence to global optima

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

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

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

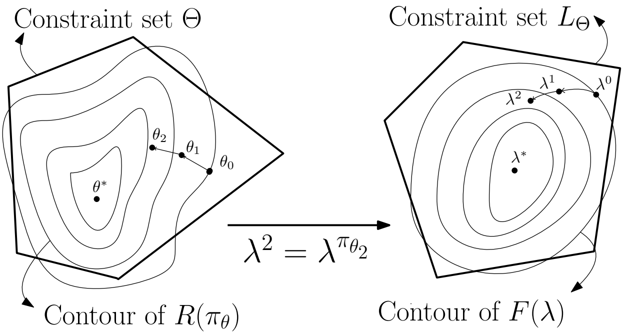


Rate of convergence to global optima

- 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.

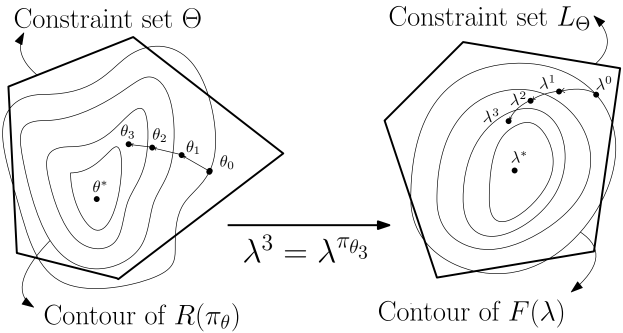


Rate of convergence to global optima

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

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

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

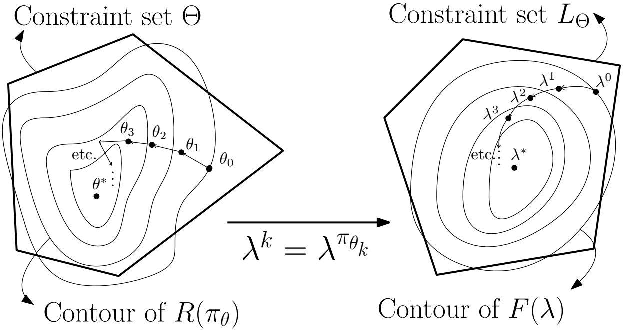


Rate of convergence to global optima

- 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.



Summary of contribution

- General RL utilities beyond cumulative reward.
- Variational Policy Gradient Theorem: estimate policy gradient for general utilities via minimax optimization.
- Global convergence of variational policy gradient updates: exploit the hidden convexity in the occupancy measure.
- State-of-the-art convergence rate.

Thank you!