Optmality Guarantees for Policy Gradient Methods

Jalaj Bhandari (with Daniel Russo)

Work done at Columbia University

February 27, 2023

Research contributions in RL

This talk ..

- "Global optimality guarantees for policy gradient methods," accepted for publication in Operations Research.

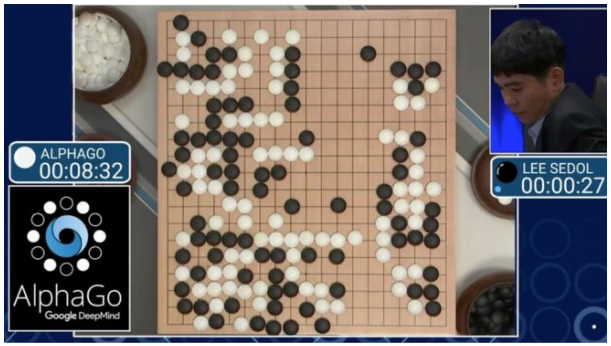
Not in this talk ..

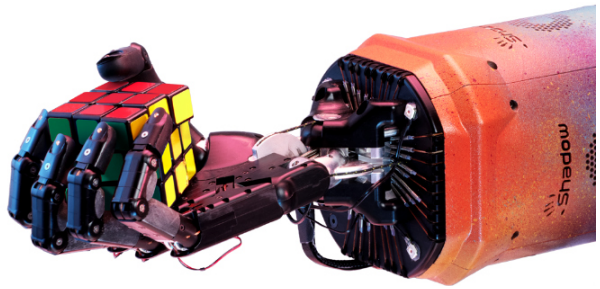
- "Linear convergence of policy gradient methods for finite MDPs," AISTATS, 2021.
- "Finite time analysis of Temporal Difference (TD) learning with linear function approximation," COLT, 2018; Operations Research, 2021.

Introduction

Surge in applications of Reinforcement learning (RL).

- End to end training using deep neural nets for state representation.
- Lots of data, powerful and (increasingly) accurate simulators.



Games, robotics, recommender systems, drug discovery etc.

Research Motivation

Foundations of Reinforcement Learning:

- Goal: Demystify RL algorithms by theoretically characterizing their behaviour under simple settings.
- Outcomes: Insights can lead to new algorithms in complex settings.

Brief introduction to RL (informally)

Find an optimal policy – a decision rule specifying (distribution over) actions which minimize the long run expected costs from each state.

Many approaches: policy iteration, policy gradient methods, Q-learning etc.

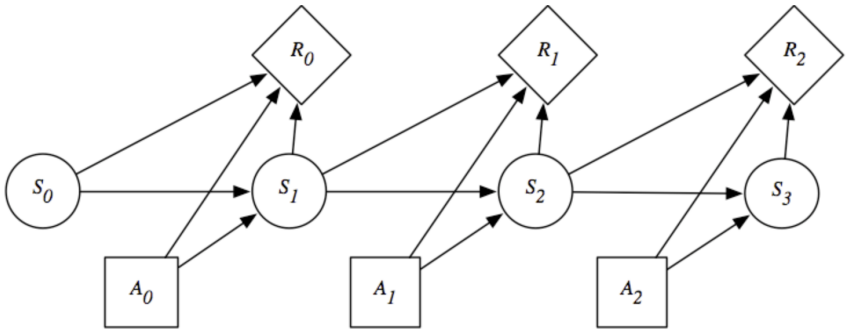
- Indirect approach: an iterative scheme

$$\pi_0 \xrightarrow{\text{Policy evaluation}} \text{evaluate}\,\pi_0 \xrightarrow{\text{Policy Improvement}} \pi_1 \to \ldots$$

- Temporal Difference learning: fundamental algorithm for policy evaluation (how good is the given policy).
- Direct approach: Policy gradient methods to improve policies using local updates.

Set-up: Markov Decision Process

Consider MDP: $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{C}, \gamma)$



- State and actions space: $\mathcal{S}, \mathcal{A}$
- Transition kernel: $\mathcal{P}(\cdot|s, a)$
- Discount factor: $\gamma \in (0, 1)$

For the purpose of this talk:

- Costs instead of rewards: $|c(s, a)| \leq c_{\max}$.
- Finite state space: $\mathcal{S} = \{s_1, s_2, \cdots, s_n\}$.
- Results extend to general state spaces under measurability assumptions, hold for all examples presented.

Set-up and background

For a policy $\pi(\cdot) : \mathcal{S} \mapsto \mathcal{A}$, define cost-to-go:

$$J_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c_\pi(s_t) \mid s_0 = s \right] \quad \forall\, s \in \mathcal{S}$$

- Here, $c_\pi(s) := c(s, \pi(s))$ and the expectation is taken over sequence of states drawn from playing policy π.

Ultimate Goal: Find an optimal policy, $\pi^*(s) = \arg\min_{\pi \in \Pi} J_\pi(s)\ \forall\, s \in \mathcal{S}$.

- Π: set of all stationary policies.

Policy gradient methods

PG objective: expected cost-to-go under some restart distribution ρ:

$$\ell(\pi) = (1-\gamma)\mathbb{E}_{s\sim\rho}[J_\pi(s)]$$

Direct policy search using first order gradient updates

- Parameterized policy class, $\Pi_\Theta = \{\pi_\theta : \theta \in \Theta\}$
- Do (stochastic) gradient descent on $\ell(\pi_\theta)$,

$$\theta_{k+1} = \theta_k - \alpha_k(\nabla\ell(\pi_{\theta_k}) + \text{noise})$$

- Compute gradients by simulation with restarts using ρ.

Policy gradient methods

Goal: optimal policy that minimizes cost-to-go for all $s \in \mathcal{S}$

$$\pi^* = \arg\min_{\pi \in \Pi} J_\pi(s),$$

Equivalently, find π^* such that $J_{\pi^*}(s) = \min_{\pi \in \Pi} J_\pi(s)$ for all $s \in \mathcal{S}$.

Assuming initial distribution $\rho(s) > 0 \, \forall \, s \in \mathcal{S}$, this is equivalent to

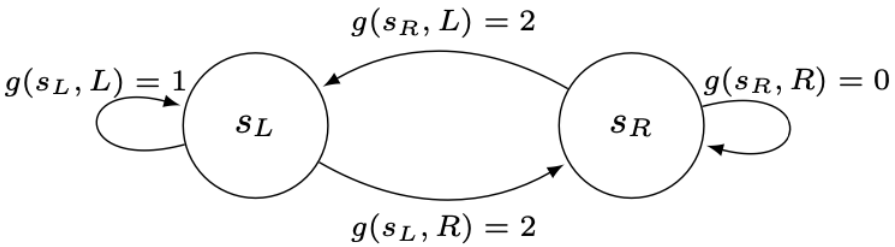
$$\pi^* = \arg\min_{\pi \in \Pi} \ell(\pi), \qquad \ell(\pi) = (1-\gamma)\mathbb{E}_{s \sim \rho}[J_\pi(s)]$$

- Avoid any challenges with exploration.
- We also assume access to exact gradient computations, $\nabla \ell(\pi)$.

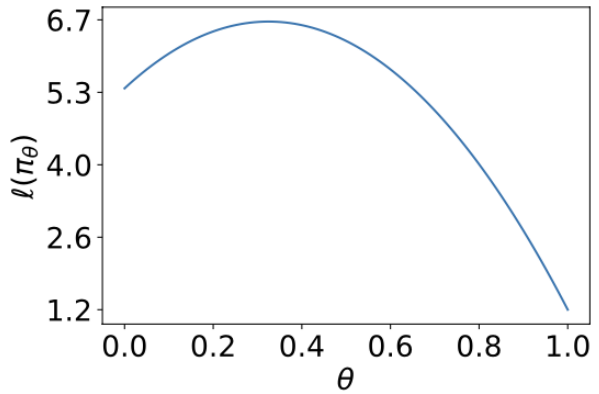
Challenge of analyzing PG methods

The policy gradient objective $\theta \mapsto \ell(\pi_\theta)$, is non-convex almost always.

- Even for the simplest case of finite state, action deterministic MDPs



(a) Two state, two action MDP



(b) $\ell(\pi_\theta)$ when $\gamma = 0.8$ and $\rho = [0.6, 0.4]$

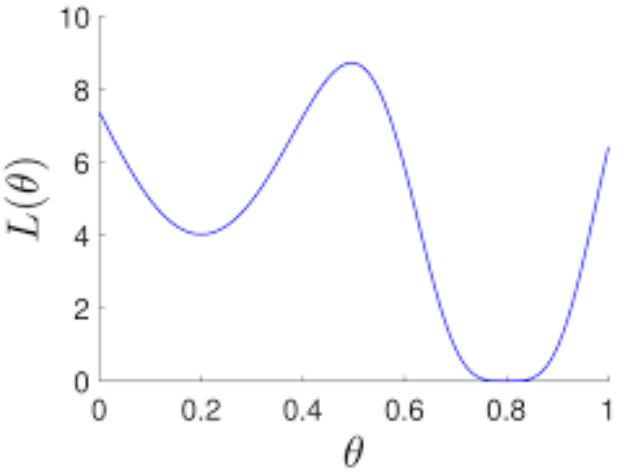
MDP with a constrained policy class, $\pi_\theta(R|s_L) = \pi_\theta(R|s_R) = \theta$.

Challenge of analyzing PG methods

Fundamental issue: cost-to-go is a multiperiod objective.

$$J_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, \pi(s_t)) \mid s_0 = s \right]$$

- Perturbing π slightly changes the distribution of states.



- First order methods for non-convex objectives asymptotically converge to stationary points.
- No guarantee on the quality of stationary points, can be arbitrarily bad.

Punchline

We study the optimization landscape of policy gradient objective, $\ell(\cdot)$.

- Sufficient exploration and exact gradients: $\theta_{k+1} = \theta_k - \alpha_k \nabla \ell(\pi_{\theta_k})$.

Main Result

We identify structural properties which guarantee:

1. Despite non-convexity, $\ell(\cdot)$ has no suboptimal stationary points. That is, any stationary point is globally optimal.
2. $\ell(\cdot)$ is "gradient dominated"

For gradient methods, (1) implies asymptotic convergence and (2) implies convergence rates with appropriate "smoothness" conditions.

Motivation

Non-convexity of policy gradient objective is a fundamental issue.

Some global optimality results despite this:

- Linear quadratic control with the class of linear policies [Fazel et. al., 2018]. Show that PG objective has no-suboptimal stationary points.
- Finite horizon inventory control with non-stationary policies [Kunnumkal and Topaloglu, 2008].

Issue: problem specific analysis. Can we do better?

Motivation

Our work:

- A general understanding of when and why we expect policy gradient methods to converge to globally optimal policies.
- Automatically applies to different problems with "structure"; atleast canonical control problems.
- Analyze landscape of the policy gradient objective rather than specific problems/algorithms

Key questions:

- What special "structure" does LQ control with the class of linear policies have? Or for that matter, finite state action mdps.

Detour: policy iteration algorithm

A step of policy optimization via policy iteration,

$$\pi_0 \xrightarrow{\text{Policy evaluation}} Q_{\pi_0}(s,a) \xrightarrow{\text{Policy Iteration}} \pi_1(s) = \arg\min_{a \in \mathcal{A}} Q_{\pi_0}(s,a)$$

A policy iteration update solves a single-period optimization problem:

$$\pi^+(s) = \arg\min_{a \in \mathcal{A}} \; Q_\pi(s,a) \quad \forall\, s \in \mathcal{S}$$

$$Q_\pi(s,a) = c(s,a) + \gamma \sum_{s'} P(s'|s,a) J_\pi(s')$$

Convergence of policy iteration (PI)

From some $\pi \in \Pi$, solve a series of single period optimization problems,

$$\pi^+(s) = \arg\min_{a \in \mathcal{A}} Q_\pi(s, a) \quad \forall\, s \in \mathcal{S}.$$

- PI updates give monotonic improvements: $J_{\pi^+} \preceq J_\pi$.
- Strict improvements for a subset of states, $J_\pi(s) - J_{\pi^+}(s) > 0$ if $\pi \neq \pi^*$ for some $s \in \mathcal{S}$.
- Policy iteration converges for linear quadratic control and tabular mdps. Special structure in the single period problems.

Key idea

From policy iteration to policy gradients:

- Single period policy iteration objective, $Q_\pi(s, \cdot)$, often has some special structure.
- Can we reason about the multi-period PG objective $\ell(\cdot)$ by exploiting the special structure in the (single period) policy iteration objective?

Illustrative example: tabular mdps

n states: $\mathcal{S} = \{1, \cdots, n\}$ and k deterministic actions.

- Stochastic policies over k actions, $\pi_\theta(s) = \theta_s \in \Delta^{k-1}$.
- Policy class: $\Pi_\Theta = \{\theta \in \mathbb{R}_+^{n \times k} : \sum_{i=1}^k \theta(s,i) = 1 \ \forall\, s\}$.

Observation 1: Π_Θ is closed under policy improvement.

- Since $\Pi_\Theta = \Pi$, the policy iteration update can be carried out in Π_Θ.
- For any $\pi_\theta \in \Pi_\Theta$ there exists a $\pi_{\theta^+} \in \Pi_\Theta$ such that,

$$\pi_{\theta^+}(s) = \arg\min_{a \in \Delta^{k-1}} Q_\pi(s,a)$$

Tabular mdps

Observation 2: Q-function is linear in a, $Q_\pi(s,a) = \sum_{i=1}^k Q_\pi(s,e_i)a_i$

- Follows as the cost and transition functions are linear:

$$c(s,a) = \sum_{i=1}^k c(s,e_i)a_i \quad \text{and} \quad P(s'|s,a) = \sum_{i=1}^k P(s'|s,e_i)a_i$$

The policy iteration problem can be solved efficiently: optimize a linear function over the simplex

$$\pi_{\theta^+}(s) = \arg\min_{a \in \Delta^{k-1}} Q_{\pi_\theta}(s,a)$$

An intuitive argument

Claim: $\theta \mapsto \ell(\theta)$ has no suboptimal stationary points.

Proof: Let $\pi_\theta = \theta$ be a suboptimal stationary point and $\pi_{\theta^+} = \theta^+$ be the corresponding policy iteration update. Set $\theta^\alpha = (1-\alpha)\theta + \alpha\theta^+$. Then,

$$J_{\pi_{\theta^\alpha}}(s) - J_{\pi_\theta}(s) \leq \underbrace{\left(\arg\min_{a \in \Delta^{k-1}} Q_{\pi_\theta}(s,a) - Q_{\pi_\theta}(s, \pi_\theta(s)) \right)}_{\text{one step policy improvement}} \leq 0$$

- Strict inequality for some state s as π_θ is a suboptimal policy (PI gives strict improvement).
- Moving toward the PI update forms a descent direction, $\frac{d}{d\alpha}\ell(\theta^\theta) < 0$

Completing the argument

Asymptotic convergence:

- Projected gradient descent applied to smooth objective converges asymptotically to stationary points.

This argument leverages:

- Convexity of the policy class and convexity of the single period policy iteration objective (Q-function).
- General conditions in the paper relax these.

Linear Quadratic control

$$\begin{aligned} \text{Minimize} \quad & \sum_{t=0}^{\infty} \gamma^t \left(a_t^\top R a_t + s_t^\top C s_t \right) \\ \text{Subject to} \quad & s_{t+1} = A s_t + B a_t, \quad s_0 \sim \rho, \quad R, C \succ 0. \end{aligned}$$

- Quadratic costs and deterministic linear dynamics.
- Some headache due to stability issues. Policy π_θ is stable if the corresponding cost-to-go is finite for all initial states.

Special structure in LQ control

A stable linear policy $\pi_\theta(s) = \theta s$ for some $\theta \in \mathcal{R}^{k \times n}$ is optimal.

Policy iteration converges to the optimal stable linear policy. The single period policy improvement problem has a special structure.

For any (stable) linear policy, π_0,

- $Q_{\pi_0}(s, a)$ is convex quadratic in a. PI problem can be solved optimally.
- $\Pi_\Theta = \{\pi_\theta \,|\, \pi_\theta(s) = \theta s,\ \theta \in \mathcal{R}^{k \times n}\}$ is closed under PI.

$$\theta_1 = \arg\min_\theta Q_{\pi_0}(s, \theta s) \in \Pi_\Theta$$

- The improved policy π_1 is also a (stable) linear policy; $J_{\pi_1} \preceq J_{\pi_0}$.

General results: weighted PI objective

Consider the following weighted PI objective. For $\pi, \pi' \in \Pi_\Theta$,

$$\mathcal{B}(\pi' \mid \eta_\pi, J_\pi) = \mathbb{E}_{s\sim\eta_\pi}\left[Q_\pi(s, \pi'(s))\right]$$

- Scalarized version under state distribution $\eta_\pi = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \rho P_\pi^t$.
- Here $\rho P_\pi^t = \mathcal{P}_\pi(s_t = s | s_0 \sim \rho)$ denotes state distribution at time t.
- Note that $\eta_\pi(s) \geq \rho(s) > 0$ by assumption, so

$$\pi^+(s) = \arg\min_{\pi'\in\Pi} Q_\pi(s, \pi'(s)) \iff \pi^+ = \arg\min_{\pi'\in\Pi} \mathcal{B}(\pi' | \eta_\pi, J_\pi)$$

Key Insight: a policy gradient theorem

Policy gradient theorem: Assuming differentiability of $\ell(\cdot)$,

$$\nabla \ell(\pi) = \nabla_{\pi'} \mathcal{B}(\pi' \mid \eta_\pi, J_\pi)\Big|_{\pi'=\pi}$$

PG methods take a gradient step w.r.t. the weighted PI objective instead of solving the PI problem to optimality.

Informal argument of no suboptimal stationary points

- Step 1: PG theorem implies that a stationary point π of $\ell(\cdot)$ is also a stationary point of $\pi' \mapsto \mathcal{B}(\pi' \mid \eta_\pi, J_\pi)$.

- Step 2: If $\pi' \mapsto \mathcal{B}(\pi' | \eta_\pi, J_\pi)$ is a convex function, i.e. has no suboptimal stationary points, and Π_Θ is closed under PI

$$\pi = \arg\min_{\pi' \in \Pi_\Theta} \mathcal{B}(\pi' | \eta_\pi, J_\pi) = \arg\min_{\pi' \in \Pi} \mathcal{B}(\pi' | \eta_\pi, J_\pi)$$

- Step 3: No progress via policy iteration $\implies$ optimal policy.

Formal conditions

Assume some regularity conditions to ensure that $\nabla \ell(\pi)$ exists.

Condition 1 (Improved policy lies in the policy class): For each $\pi \in \Pi_\Theta$, there exists $\pi^+ \in \Pi_\Theta$ such that $\pi^+ = \arg\min_{\pi' \in \Pi} \mathcal{B}(\pi' | \eta_\pi, J_\pi)$.

Condition 2: (Stationary points of the Bellman objective) For all $\pi \in \Pi_\Theta$, the function $\pi' \mapsto \mathcal{B}(\pi' \mid \eta_\pi, J_\pi)$ has no sub-optimal stationary points.

Theorem: Under Conditions 1 and 2 (and some regularity conditions), $\pi \in \Pi_\Theta$ is a stationary point of $\ell(\cdot)$ if and only if $J_\pi = J^*$.

Convergence rate for non-convex optimization: Idea of gradient dominance

Polyak-Lojasiewicz (PL) inequality for unconstrained non-convex functions,

$$f(x) - \min_{x' \in \mathcal{R}^d} f(x') \leq \frac{1}{2\mu} \|\nabla f(x)\|_2^2, \text{ for some } \mu > 0$$

- Optimality gap at point x upper bounded by gradient norm at x.
- Implies no sub-optimal stationary points. But weaker than convexity.
- For smooth objectives, gradient dominance implies linear convergence with gradient descent.

(c, μ)-gradient dominance

A function f is (c, μ)-gradient dominated over $\mathcal{X} \subset \mathcal{R}^d$ for $c > 0$ and $\mu \geq 0$ if for all $x \in \mathcal{X}$,

$$\min_{x' \in \mathcal{X}} f(x') \geq f(x) + \min_{x' \in \mathcal{X}} \left[c \left\langle \nabla f(x), x' - x \right\rangle + \frac{\mu}{2} \|x - x'\|_2^2 \right]$$

Recall, strongly convex functions,

$$f(x') \geq f(x) + \left\langle \nabla f(x), x' - x \right\rangle + \frac{\mu}{2} \|x - x'\|_2^2$$

- Cvx functions are $(1, 0)$ g.d.; μ-strongly cvx functions are $(1, \mu)$ g.d.
- $\mathcal{O}(1/\sqrt{T})$ convergence for projected gradient descent with smoothness conditions.

Result - Gradient dominance of $\ell(\cdot)$

Condition 3 (Gradient dominance of weighted policy iteration objective)
For any $\pi \in \Pi_\Theta$, the function $\theta \mapsto \mathcal{B}(\theta \mid \eta_\pi, J_\pi)$ is (c, μ) gradient dominated over Θ.

Theorem: Under Conditions 1, and 3 and some regularity conditions, $\ell(\cdot)$ is $\left(\frac{1-\gamma}{\kappa_\rho} \cdot c\,,\ \frac{1-\gamma}{\kappa_\rho} \cdot \mu\right)$ gradient dominated.

- Concentrability coefficient: κ_ρ measures the "effectiveness" of the initial distribution. See paper for details.

Extensions and other results

See paper for more details on different examples, gradient dominance and concentrability coefficient κ_ρ.

Extension: The idea of approximate closure

$$\left| \min_{\pi' \in \Pi} \mathcal{B}(\pi' | \eta_\pi, J_\pi) - \min_{\pi' \in \Pi} \mathcal{B}(\pi' | \eta_\pi, J_\pi) \right| \leq \epsilon$$

- Approximation error of performing the PI update in Π_Θ is bounded.
- Any stationary point of $\ell(\cdot)$ is nearly optimal.

Extension: state aggregation

We understand tabular MDPs very well. What about continuous state space $\mathcal{S}$? One way:

- Partition $\mathcal{S}$ into m aggregate states, $\mathcal{S}_1, \cdots, \mathcal{S}_m$.
- Stochastic policies over subsets such that: $\pi(s, i) = \pi(\mathcal{S}_j, i)\ \forall\, s \in \mathcal{S}_j$.
- "Approximate" policy closure:

$$\left| \min_{\pi' \in \Pi_{\mathcal{S}}} \mathcal{B}(\pi' \mid \eta_\pi, J_\pi) - \min_{\pi' \in \Pi} \mathcal{B}(\pi' \mid \eta_\pi, J_\pi) \right| \leq \epsilon$$

- See paper for a characterization of ϵ.
- How do we find such aggregate states $\{\mathcal{S}_1, \cdots, \mathcal{S}_m\}$.