Policy Gradient Methods in Markov Decision Processes

Ziniu Li
ziniuli@link.cuhk.edu.cn

The Chinese University of Hong Kong, Shenzhen, Shenzhen, China

April 9, 2021

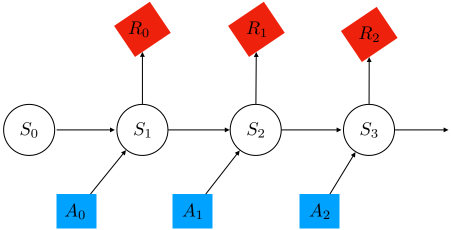
Mainly based on the COLT 2020 paper:
Agarwal, Alekh, et al. "Optimality and approximation with policy gradient methods in markov decision processes." Annual Conference on Learning Theory, 2020.

Outline

Outline

Markov Decision Process

- An infinite-horizon discounted Markov Decision Process (MDP) [Puterman, 2014] is described by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma, \rho)$:
  - $\mathcal{S}$ and $\mathcal{A}$ are the finite state and action space, respectively.
  - $p(s'|s, a)$ is the transition probability matrix.
  - $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is the deterministic reward function.
  - $\gamma \in (0, 1)$ is the discount factor.
  - ρ specifies the initial state distribution.



Markov Decision Process: Policy

- To interact with MDP, we need a policy π to select actions.
  - $\pi(a|s)$ determines the probability of selecting action a at state s.
- The quality of policy π is measured by state value function V^π:

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | \pi, s_0 = s\right]. \tag{1}$$

  - $V^\pi(s)$ measures the the expected long-term discounted reward when starting from state s.
  - $V^\pi(s) \in [0, \frac{1}{1-\gamma}]$ by definition.
- To take the initial state distribution into account, we define

$$V^\pi(\rho) = \mathbb{E}_{s_0 \sim \rho}\left[V^\pi(s_0)\right]. \tag{2}$$

Markov Decision Process: Value Function

- Sometimes, it is more convenient to introduce state-action value function Q^π:

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | \pi, s_0 = s, a_0 = a\right]. \tag{3}$$

  - $Q^\pi(s,a)$ measures the the expected long-term discounted reward when starting from state s with action a.
  - $V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s,a)]$ by definition.

- To further qualify the benefits of selecting an action a, we introduce advantage function $A^\pi(s,a)$:

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s). \tag{4}$$

♠ : $\sum_{a\in\mathcal{A}} \pi(a|s) A^\pi(s,a) = 0.$

Markov Decision Process: Discounted Stationary Distribution

- To facilitate later analysis, we introduce discounted stationary distribution d^{π}:

$$d_{s_0}^{\pi}(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s | \pi, s_0). \tag{5}$$

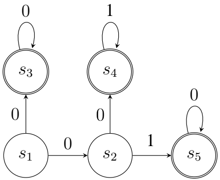
♠ : $d_{s_0}^{\pi}(s)$ measures the discounting probability to visit s starting from the initial state s_0.

- To take the initial state distribution into account, we define d_{ρ}^{π} as

$$d_{\rho}^{\pi}(s) = \mathbb{E}_{s_0 \sim \rho} \left[d_{s_0}^{\pi}(s) \right]. \tag{6}$$

Markov Decision Process: Example

- Consider the following MDP example: a_1 : "up"; a_2 : "right".



$$\pi(a_2|s_1) = 1;$$

$$\pi(s_2|a_1) = 0.5,\ \pi(s_2|a_2) = 0.5.$$

$$d_\rho^\pi(\cdot) = (1-\gamma)\cdot\left(1, \gamma, 0, 0.5\tfrac{\gamma^2}{1-\gamma}, 0.5\tfrac{\gamma^2}{1-\gamma}\right).$$

$$V^\pi(s_1) = 0 + 0.5\,[\gamma \times 1] + 0.5\left[\gamma^2 \times 1 + \gamma^3 \times 1 + \cdots\right] = \frac{0.5\gamma}{1-\gamma},$$

$$Q^\pi(s_2, a_1) = 1, \quad Q^\pi(s_2, a_2) = \frac{\gamma}{1-\gamma}, \quad V^\pi(s_2) = \frac{0.5}{1-\gamma},$$

$$A^\pi(s_2, a_1) = 1 - \frac{0.5}{1-\gamma}, \quad A^\pi(s_2, a_2) = \frac{\gamma - 0.5}{1-\gamma}.$$

♠ : at state s_2, a_1 has a larger advantage if $\gamma > 0.5$.

Outline

Optimization Theory: Nonconvexity & Smothness

- We mainly focus on gradient-based optimization methods, which are shown to converge to some stationary point for any differentiable function.
- For differentiable functions, we can categorise them according to convexity.
  - for convex functions, all stationary points are global;
  - for nonconvex functions, if there is no saddle point, the stationary point is a local minima.
- To better qualify the convergence rate, we often assume the gradient of differentiable is β-Lipschitz continuous (sometimes, we call it β-smooth).
  - specifically, β-smooth means that

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|, \quad \forall x, y \in \mathbb{R}^n. \tag{7}$$

  - by simple calculus, β-smooth implies the following upper bound

$$f(y) \leq f(x) - \langle \nabla f(x), y - x \rangle + 0.5\beta \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n \tag{8}$$

Optimization Theory: Nonconvexity & Smothness

- Consider the simplest algorithm: gradient descend.
$$x^{t+1} = x^t - \eta \nabla f(x^t). \tag{9}$$
- β-smooth implies:
$$\begin{aligned} f(x^{t+1}) &\le f(x^t) - \langle \nabla f(x^t), x^{t+1} - x^t \rangle + 0.5\beta \left\| x^{t+1} - x^t \right\|^2 \\ &= f(x^t) - \eta \left(1 - 0.5\beta\eta\right) \left\| \nabla f(x^t) \right\|^2 . \end{aligned}$$
♠ : if η is small enough (i.e., $\eta < 2\beta^{-1}$), we have $f(x^{t+1}) \le f(x^t)$, which is called the descend property in smooth optimization.
- Considering the stepsize $\eta = \frac{1}{\beta}$, we obtain that
$$\left\| \nabla f(x^t) \right\|^2 \le 2\beta \left(f(x^t) - f(x^{t+1}) \right). \tag{10}$$

Optimization Theory: Nonconvexity & Smothness

- Summing up (10) over $t = 0, 1, \cdots, T-1$, we obtain that

$$\begin{aligned}
\sum_{t=0}^{T-1} \left\|\nabla f(x^t)\right\|^2 &\leq 2\beta\left(f(x^0) - f(x^T)\right) \\
\implies \quad \min_{t=0,\cdots,T-1} \left\|\nabla f(x^t)\right\|^2 &\leq \frac{2\beta\left(f(x^0) - f(x^T)\right)}{T} \\
\implies \quad \min_{t=0,\cdots,T-1} \left\|\nabla f(x^t)\right\| &\leq \sqrt{\frac{2\beta\left(f(x^0) - f(x^T)\right)}{T}}.
\end{aligned} \tag{11}$$

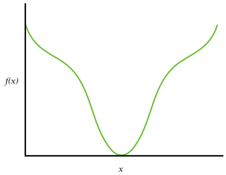
- To summarize, for nonconvex smooth optimization, the speed to find a stationary point /local minima is in order of $\mathcal{O}(\sqrt{\beta/T})$.

Optimization Theory: Polyak-Lojasiewicz (PL) condition:

- The β-smooth assumption leads to a local convergence result and we cannot get a global convergence in general nonconvex optimization
- However, Polyak-Lojasiewicz (PL) condition [Polyak, 1963, Lojasiewicz, 1963] implies gradient domination and global convergence:

$$\|\nabla f(x)\|^2 \geq \mu \left(f(x) - \min_x f(x) \right),$$

♠ : all stationary points are global miniziers!



$f(x) = x^2 + \sin^2(x)$, which is nonconvex but satisfies the PL condition.

Optimization Theory: Short Summary

- Two properties are important for convergence analysis: smoothness and regularity conditions.
- Smoothness (such as gradient Lipschitz continuous) ensures "descend" trend.
- Regularity conditions (such as Polyak-Lojasiewicz (PL) condition) ensures the global convergence.
- To establish the global convergence of policy gradient methods, we need to first identify these properties, which will be shown later.

Optimization Theory: Constrained Optimization

- In previous, we consider the unconstrained optimization $\mathcal{X} = \mathbb{R}^n$. Here, we briefly review the optimality condition for constrained optimization.
- Consider a convex set $\mathcal{X}$, a direct sufficient first-order optimality condition is
$$d^\top \nabla f(x^*) \geq 0, \quad \forall d \in \mathcal{R}_{\mathcal{X}}(x^*),$$
where $\mathcal{R}_{\mathcal{X}}(x^*) = \{d \in \mathbb{R}^n : \exists t^* > 0 \text{ such that } x + td \in \mathcal{X} \text{ for all } t \in [0, t^*)\}$.
- Note that equivalently, we have
$$(x - x^*)^\top \nabla f(x^*) \geq 0, \quad \forall x \in \{x^*\} + t\mathcal{R}_{\mathcal{X}}(x^*).$$
- Therefore, gradient domination in PL condition need to be revised in this setting.
  - Specific form is given in later (see Lemma 3).

Outline

Outline

Policy Gradient Theorem

Theorem 1 (Policy Gradient Theorem).

Consider policy π is parameterized by θ, then we have for any state $s_0 \in \mathcal{S}$,

$$\nabla_\theta V^{\pi_\theta}(s_0) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \mathbb{E}\left[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a)\right]. \tag{12}$$

In addition, if this parameterization satisfies the simplex constraint, that is $\sum_{a \in \mathcal{A}} \pi_\theta(a|s) = 1$, we further have:

$$\nabla_\theta V^{\pi_\theta}(s_0) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \mathbb{E}\left[\nabla_\theta \log \pi_\theta(a|s) A^{\pi_\theta}(s,a)\right]. \tag{13}$$

♠ : As long as we know exact Q^{π_θ} (or A^{π_θ}) and d^{π_θ}, policy gradient is available.

♠ : This assumption is very strong in practice, where it is impossible to evaluate a policy π perfectly (instead, people use Monte Carlo estimation).

Outline

Policy Gradient Methods: Direct Parameterization

- It is natural to build up a table $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, in which $\theta_{sa} = \pi(a|s)$ is the probability of selecting action a at state s. We call this direct parameterization.
  - To satisfy the probability simplex condition, we require $\sum_{a\in\mathcal{A}} \theta_{sa} = 1$ for every state s.
- Note that the policy gradient in (12) or (13) does not hold for this parameterization.
  - This is because $\sum_a \nabla_\theta \pi_\theta(a|s) = 0$ is not explicitly maintained by the direct parameterization.
- By inspecting the proof of Theorem 1, we get the following expression for direct parameterization:

$$\nabla_\theta V^{\pi_\theta}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi_{s_0}}\left[\sum_{a\in\mathcal{A}} \nabla \pi_\theta(a|s) Q^\pi(s,a)\right].$$

- Since θ_{sa} and $\pi(a|s)$ is one-to-one, we may write the gradient as ∇_π instead of ∇_θ. Then, focusing on the single $\pi(a|s)$ element, we get

$$\frac{\partial V^\pi(\mu)}{\partial \pi(a|s)} = \frac{1}{1-\gamma} d^\pi_\mu(s) Q^\pi(s,a). \tag{14}$$

Policy Gradient Methods: Softmax Parameterization

- To avoid the probability simplex, we can use the so-called softmax parameterization:

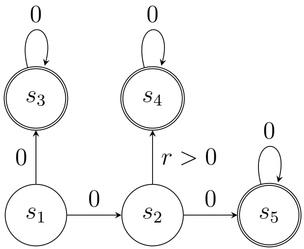
$$\pi(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}. \tag{15}$$

- Advantage: the associated optimization problem is unconstrained.
- Disadvantage: it cannot approximate the deterministic policy in finite regime.

Outline

Nonconcavity of Policy Gradient Optimization

Lemma 1.
There is an MDP $\mathcal{M}$ (see Figure 2) such that the optimization problem is not concave for both direct parameterization and softmax parameterization.



A simple MDP example corresponding to Lemma 1. For this MDP, both direct parameterization and softmax parameterization yield a nonconcave optimization problem. Figure from [Agarwal et al., 2020].

Nonconcavity of Policy Gradient Optimization

- To understand the idea in Lemma 1, let us consider a simple function:

$$f(x,y) = xy, \quad \nabla f(x) = \begin{bmatrix} y \\ x \end{bmatrix}, \quad \nabla^2 f(x,y) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{16}$$

  ♠ : this function is convex/concave w.r.t x or y, but is neither convex or concave w.r.t. (x,y).

- Informally,

$$\text{expected return} = \sum_{\text{trajectory}} \mathbb{P}(\text{trajectory}) \times R(\text{trajectory}).$$

- We see that $\mathbb{P}(\text{trajectory}) = \prod_{t=0}^{\infty} \pi(a_t|s_t)p(s_{t+1}|s_t,a_t)$ could have the structure in (16); therefore, we expect it is nonconcave for policy optimization.

Proof of Lemma 1

- Let a_1 be the "up" action and a_2 be "right" action in the MDP displayed in Figure 2.
- We see that only the "up" action at s_2 has positive reward. Consider the initial state as s_1, then we have
$$V^{\pi}(s_1) = \pi(a_2|s_1)\pi(a_1|s_2) \cdot r.$$
- Consider $\theta = (\theta_{a_1,s_1}, \theta_{a_2,s_1}, \theta_{a_1,s_2}, \theta_{a_2,s_2})$ with the following values:
$$\theta^{(1)} = (\log 1, \log 3, \log 3, \log 1), \quad \theta^{(2)} = (-\log 1, -\log 3, -\log 3, -\log 1).$$
- For the softmax parameterization, we have that
$$\pi^{(1)}(a_2|s_1) = \frac{3}{4}; \quad \pi^{(1)}(a_1|s_2) = \frac{3}{4}; \quad V^{(1)}(s_1) = \frac{9}{16}r;$$
$$\pi^{(2)}(a_2|s_1) = \frac{1}{4}; \quad \pi^{(2)}(a_1|s_2) = \frac{1}{4}; \quad V^{(2)}(s_1) = \frac{1}{16}r;$$

Proof of Lemma 1

- Now, consider $\theta^{(\text{mid})} = (\theta^{(1)} + \theta^{(2)})/2$,

$$\pi^{(\text{mid})}(a_2|s_1) = \frac{1}{2}; \quad \pi^{(\text{mid})}(a_1|s_2) = \frac{1}{2}; \quad V^{(\text{mid})}(s_1) = \frac{1}{4}r;$$

- We see that $V^{(1)}(s_1) + V^{(2)}(s_1) > 2V^{(\text{mid})}(s_1)$; thus, the optimization problem for softmax parameterization is not concave.
- Finally, note the above argument also holds for the direct parameterization.

Smoothness in Policy Optimization

Lemma 2 (Smothness for direct parameterization).

For all starting states s_0 *and policies* π, π'*, we have*

$$\left\| \nabla_\pi V^\pi(s_0) - \nabla_\pi V^{\pi'}(s_0) \right\|_2 \leq \frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3} \left\| \pi - \pi' \right\|_2 . \tag{17}$$

♠ : the proof is very technical and can be found in [Agarwal et al., 2020].

Gradient Domination in Policy Optimization

Lemma 3 (Gradient domination for direct parameterization).

For the direct policy parameterization, for all state distributions $\mu, \rho \in \Delta(\mathcal{S})$*, we have*

$$V^*(\rho) - V^\pi(\rho) \le \left\| \frac{d_\rho^{\pi^*}}{d_\mu^\pi} \right\|_\infty \max_{\overline{\pi}} (\overline{\pi} - \pi)^\top \nabla_\pi V^\pi(\mu) \tag{18}$$

$$\le \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \max_{\overline{\pi}} (\overline{\pi} - \pi)^\top \nabla_\pi V^\pi(\mu), \tag{19}$$

where $\max$ *is over the set of all policies, i.e.,* $\overline{\pi} \in \Delta(A)^{|\mathcal{S}|}$.

♠ : the reasoning from (18) to (19) is easy because
$d_\mu^\pi(s) := (1-\gamma) \sum_{t=0}^\infty \gamma^t \mathbb{P}(s_t = s | \pi, s_0 \sim \mu)$, which implies $d_\mu^\pi(s) \ge (1-\gamma)\mu(s)$ for all $s \in \mathcal{S}$.

Gradient Domination in Policy Optimization: Proof of Lemma 3

- The proof of Lemma 3 relies on another famous lemma.

Lemma 4 (Performance difference lemma[Kakade and Langford, 2002]).

For all policies π and π' and state s_0, we have

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi}_{s_0}} \mathbb{E}_{a \sim \pi(\cdot|s)} \left[A^{\pi'}(s,a) \right]. \tag{20}$$

- Intuitively, Lemma 4 states that the value difference between two policies equal to the expected advantages of the reference policy π' over the distribution induced by the evaluation policy π.

Gradient Domination in Policy Optimization: Proof of Lemma 3

Based on performance difference lemma (Lemma 4), we have

$$
\begin{aligned}
V^*(\rho) - V^\pi(\rho) &= \frac{1}{1-\gamma}\mathbb{E}_{s_0\sim\rho}\mathbb{E}_{s\sim d_{s_0}^{\pi^*}}\mathbb{E}_{a\sim\pi^*(\cdot|s)}\left[A^\pi(s,a)\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi^*}}\mathbb{E}_{a\sim\pi^*(\cdot|s)}\left[A^\pi(s,a)\right] \\
&\leq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi^*}}\mathbb{E}_{a\sim\pi^*(\cdot|s)}\left[\max_{\bar{a}} A^\pi(s,\bar{a})\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi^*}}\left[\max_{\bar{a}} A^\pi(s,\bar{a})\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi}}\left[\frac{d_\rho^{\pi^*}(s)}{d_\mu^\pi(s)}\max_{\bar{a}} A^\pi(s,\bar{a})\right] \\
&\leq \frac{1}{1-\gamma}\left(\max_s \frac{d_\rho^{\pi^*}(s)}{d_\mu^\pi(s)}\right)\mathbb{E}_{s\sim d_\mu^{\pi}}\left[\max_{\bar{a}} A^\pi(s,\bar{a})\right].
\end{aligned}
$$

Gradient Domination in Policy Optimization: Proof of Lemma 3

Next, connect $\max_{\bar{a}} A^\pi(s, \bar{a})$ with policy gradient:

$$\begin{aligned}
\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^\pi}\left[\max_{\bar{a}} A^\pi(s,\bar{a})\right] &= \max_{\overline{\pi}\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \sum_{s,a} \frac{d_\mu^\pi(s)}{1-\gamma}\overline{\pi}(a|s)A^\pi(s,a) \\
&= \max_{\overline{\pi}\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \sum_{s,a} \frac{d_\mu^\pi(s)}{1-\gamma}\left(\overline{\pi}(a|s)-\pi(a|s)\right)A^\pi(s,a) \\
&= \max_{\overline{\pi}\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \sum_{s,a} \frac{d_\mu^\pi(s)}{1-\gamma}\left(\overline{\pi}(a|s)-\pi(a|s)\right)Q^\pi(s,a) \\
&= \max_{\overline{\pi}\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \left(\overline{\pi}-\pi\right)^\top \nabla_\pi V^\pi(\mu),
\end{aligned}$$

Gradient Domination in Policy Optimization: Remark

Combing the above inequalities, we obtain the gradient domination theorem:

$$\begin{aligned} V^*(\rho) - V^\pi(\rho) &\le \left\| \frac{d_\rho^{\pi^*}}{d_\mu^\pi} \right\|_\infty \max_{\overline{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\overline{\pi} - \pi)^\top \nabla_\pi V^\pi(\mu) \\ &\le \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \max_{\overline{\pi}} (\overline{\pi} - \pi)^\top \nabla_\pi V^\pi(\mu). \end{aligned}$$

- We quickly get that $\max_{\overline{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\overline{\pi} - \pi)^\top \nabla_\pi V^\pi(\mu) = 0 \Longrightarrow V^\pi(\rho) = V^*(\rho)$.
- The optimization distribution is allowed to be different from the initial state distribution ρ, and the distribution mismatch coefficient $\left\| d_\rho^{\pi^*} / \mu \right\|_\infty$ captures the exploration difficulty.
  - Note that $\left\| d_\rho^{\pi^*} / \mu \right\|_\infty$ could be exponentially large for hard exploration problems (see [Agarwal et al., 2020, Section 4.3].

Outline

Outline

Projected Gradient Ascend

- We consider a very simple algorithm: projected gradient ascend (PGA):
$$\pi^{(t+1)} = \mathcal{P}_{\Delta(\mathcal{A})^{|\mathcal{S}|}}\left(\pi^t + \eta \nabla_\pi V^{(t)}(\mu)\right), \tag{21}$$
where $\mathcal{P}_{\Delta(\mathcal{A})^{|\mathcal{S}|}}$ denotes the projection on the probability simplex $\Delta(\mathcal{A})^{|\mathcal{S}|}$ in terms of the Euclidean norm, that is,
$$\mathcal{P}_{\Delta(\mathcal{A})^{|\mathcal{S}|}}(\pi) \in \underset{\overline{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}}{\operatorname{argmin}} \|\pi - \overline{\pi}\|_2^2 .$$

Projection onto a simplex
Set $C := \{x \in \mathbb{R}^n : x \geq 0, \sum_{i=1}^n x_i = \eta\}$, $\eta > 0$, then $$[\mathcal{P}_C(x)]_i = \max\{0, x_i - \tau\}.$$ The factor $\tau \in \mathbb{R}$ is determined as follows: let $y = x^\downarrow$ be a sorted copy of x with $y_1 = x_1^\downarrow \geq y_2 = x_2^\downarrow \geq \ldots \geq y_n = x_n^\downarrow$. Calculate $\tau_i = [\sum_{k=1}^i y_i - \eta]/i$ and $q = \max\{j \in \{1, \ldots, n\} : y_j > \tau_j\}$ and set $\tau = \tau_q$.

Projection onto a simplex. Figure from DDA6010, Fall 2020 by Andre Milzarek at CUHKSZ.

Projected Gradient Ascend

Theorem 2 (Global convergence of projected gradient ascent for direct parameterization).

For any initial state distribution ρ*, the projected gradient ascent algorithm (21) with stepsize* $\eta = \frac{(1-\gamma)^3}{2\gamma|\mathcal{A}|}$ *satisfies*

$$\min_{t<T} \left\{ V^*(\rho) - V^{(t)}(\rho) \right\} \leq \varepsilon, \quad \textit{whenever} \quad T \geq \frac{64\gamma|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^6 \varepsilon^2} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty^2 . \tag{22}$$

♠ : This result should not be surprised since we know that the gradient norm of PGA converges in $\mathcal{O}(\sqrt{\beta/T})$ with $\beta = 2\gamma|\mathcal{A}|(1-\gamma)^{-3}$ (see Lemma 2); see the detailed proof for the additional $(1-\gamma)^3\sqrt{|\mathcal{S}|}$ term.

Outline

Proof of Theorem 2

Recall that the proof is almost done by the following steps:

(1) Show that the norm of gradient mapping G^η (i.e., the generalized gradient norm) diminishes in $\mathcal{O}(\sqrt{\beta/T})$ (i.e., the generalized result in (11) for constrained optimization).

$$G^\eta = \frac{1}{\eta}\left(\mathcal{P}_{\Delta(\mathcal{A})^{|\mathcal{S}|}}\left(\pi + \eta \nabla_\pi V^\pi(\mu)\right) - \pi\right), \tag{23}$$

(2) (New) Show that a small $\|G^\eta\|$ implies the optimality condition, that is,

$$\|G^\eta\| \leq \varepsilon \quad \Longrightarrow \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \left(\bar{\pi} - \pi\right)^\top \nabla_\pi V^\pi(\mu) \leq \mathcal{O}\left(\varepsilon\right).$$

(3) Apply the gradient domination lemma (Lemma 3) to show that $V^*(\rho) - V^\pi(\rho)$ is small conditioned on $\max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \left(\bar{\pi} - \pi\right)^\top \nabla_\pi V^\pi(\mu)$ is small.

Proof of Theorem 2: Step (1)

This step directly follows the classical convergence result of project gradient result on nonconvex optimization problems.

Lemma 5.

For a β-smooth function, with stepsize $\eta = 1/\beta$, the projected gradient descend algorithms satisfies:

$$\min_{t=0,\cdots,T-1} \left\| G^\eta(x^t) \right\| \leq \sqrt{\frac{2\beta \left(f(x^T) - \min_x f(x) \right)}{T}},$$

where $G^\eta(x)$ is called gradient mapping:

$$G^\eta(x) = \frac{1}{\eta} \left(x - \mathcal{P}_{\mathcal{X}} \left(x - \eta \nabla f(x) \right) \right).$$

See [Beck, 2017, Theorem 10.15] for the proof.

Proof of Theorem 2: Step (2)

Step (2) uses the following useful proposition.

Proposition 1.

Let $V^{\pi}(\mu)$ *be* β*-smooth in* π*. Define the gradient mapping*

$$G^{\eta} = \frac{1}{\eta}\left(\mathcal{P}_{\Delta(\mathcal{A})^{|\mathcal{S}|}}\left(\pi + \eta \nabla_{\pi} V^{\pi}(\mu)\right) - \pi\right), \tag{24}$$

and the update rule for the projected gradient is $\pi^{+} = \pi + \eta G^{\eta}$*. If* $\|G^{\eta}\|_2 \leq \varepsilon$*, then*

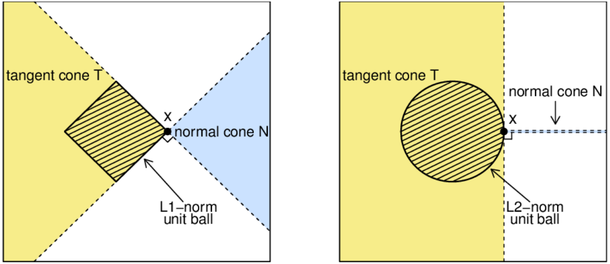
$$\max_{\pi^{+}+\delta \in \Delta(\mathcal{A})^{|\mathcal{S}|}, \|\delta\|_2 \leq 1} \delta^{\top} \nabla_{\pi} V^{\pi^{+}}(\mu) \leq \varepsilon(\eta\beta + 1). \tag{25}$$

Proof of Theorem 2: Step (2)

Proof: By [Ghadimi and Lan, 2016, Lemma 3], we have

$$\nabla_\pi V^{\pi^+}(\mu) \in N_{\Delta(\mathcal{A})^{|\mathcal{S}|}}(\pi^+) + \varepsilon(\eta\beta + 1)B_2,$$

where $N_{\Delta(\mathcal{A})^{|\mathcal{S}|}}(\pi^+)$ is the normal cone of the product simplex $\Delta(\mathcal{A})^{|\mathcal{S}|}$ and B_2 is the unit ball. Note that δ is in the tangent cone, we quickly get the desired result.



Tangent cone and normal cone. Figure from [Foygel and Mackey, 2014].

Proof of Theorem 2: Step (2)

- From step (1) (Lemma 5) and Proposition 1, we have

$$\min_{t=0,\cdots,T-1} \max_{\pi^{(t)}+\delta\in\Delta(\mathcal{A})^{|\mathcal{S}|},\|\delta\|_2\le 1} \delta^\top \nabla_\pi V^{\pi^{(t)}}(\mu) \le (\eta\beta+1)\sqrt{\frac{2\beta\left(V^*(\mu)-V^{(0)}(\mu)\right)}{T}}. \tag{26}$$

  ♠ : by choosing $\eta = 1/\beta$, the coefficient in RHS of (26) becomes 2.

- Then, we remains to connect $\max_{\pi^{(t)}+\delta\in\Delta(\mathcal{A})^{|\mathcal{S}|},\|\delta\|_2\le 1} \delta^\top \nabla_\pi V^{\pi^{(t)}}(\mu)$ with $\max_{\overline{\pi}\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \left(\overline{\pi}-\pi\right)^\top \nabla_\pi V^{\pi}(\mu)$ in the gradient domination lemma.

Proof of Theorem 2: Step (2)

Observe that

$$
\begin{aligned}
\max_{\overline{\pi}\in\Delta(\mathcal{A})^{|\mathcal{S}|}} (\overline{\pi}-\pi)^\top \nabla_\pi V^\pi(\mu) &= 2\sqrt{|\mathcal{S}|} \max_{\overline{\pi}\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \frac{1}{2\sqrt{|\mathcal{S}|}} (\overline{\pi}-\pi)^\top \nabla_\pi V^\pi(\mu) \\
&\overset{\delta:=\overline{\pi}-\pi}{=\!=} 2\sqrt{|\mathcal{S}|} \max_{\pi+\delta\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \frac{1}{2\sqrt{|\mathcal{S}|}} \delta^\top \nabla_\pi V^\pi(\mu) \\
&\overset{(1)}{\leq} 2\sqrt{|\mathcal{S}|} \max_{\pi+\delta\in\Delta(\mathcal{A})^{|\mathcal{S}|}, \|\delta\|_2\leq 2\sqrt{|\mathcal{S}|}} \frac{1}{2\sqrt{|\mathcal{S}|}} \delta^\top \nabla_\pi V^\pi(\mu) \\
&\overset{(2)}{=} 2\sqrt{|\mathcal{S}|} \max_{\pi+\delta\in\Delta(\mathcal{A})^{|\mathcal{S}|}, \|\delta\|_2\leq 1} \delta^\top \nabla_\pi V^\pi(\mu), \qquad (27)
\end{aligned}
$$

where (1) is based on the fact that $\|\delta\| = \|\overline{\pi}-\pi\|_2 \leq 2\sqrt{|\mathcal{S}|}$, and (2) follows that if $\pi+\delta\in\Delta(\mathcal{A})^{|\mathcal{S}|}$, then $\pi+c\delta\in\Delta(\mathcal{A})^{|\mathcal{S}|}$ for some constant c and importantly if δ attains the maximal value, so $c\delta$ does for $c>0$.

Proof of Theorem 2: Step (3)

Using the gradient domination lemma (Lemma 3) with (26) and (27), we have

$$
\begin{aligned}
\min_{t=0,\cdots,T-1} V^*(\rho) - V^{(t)}(\rho) &\overset{(27)}{\leq} \min_{t=0,\cdots,T-1} \frac{2\sqrt{|\mathcal{S}|}}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \max_{\pi^{(t)}+\delta\in\Delta(\mathcal{A})^{|\mathcal{S}|}, \|\delta\|_2\leq 1} \delta^\top \nabla_\pi V^{\pi^{(t)}}(\mu) \\
&\overset{(26)}{\leq} \frac{2\sqrt{|\mathcal{S}|}}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty (\eta\beta+1)\sqrt{\frac{2\beta\left(V^*(\mu)-V^{(0)}(\mu)\right)}{T}} \\
&= \frac{4\sqrt{|\mathcal{S}|}}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \sqrt{\frac{2\gamma|\mathcal{A}|\left(V^*(\mu)-V^{(0)}(\mu)\right)}{(1-\gamma)^3 T}},
\end{aligned}
$$

where the last step follows the choice of η and the definition of β. Note that $V^*(\mu) - V^{(0)}(\mu) \leq 2(1-\gamma)^{-1}$; so letting $T \geq \frac{64\gamma|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^6\varepsilon^2} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty^2$, we can ensure RHS is smaller than ε.

Outline

Summary

- From the view of optimization, smoothness and regularity conditions provide necessary properties to build global convergence rate.
- We verify the smoothness in Lemma 2 (though actually we do not).
- We verify the gradient domination condition in Lemma 3.
- The convergence rate of projected gradient ascent is obtained in Theorem 2.

Extension: Other RL Problems

- The same routine (e.g., smoothness and gradient domination verification) in linear quadratic regulator (LQR).
  - https://antonxue.github.io/sketches/antonxue-ese680-final-report.pdf
- Extension to generative adversarial imitation learning [Cai et al., 2019, Zhang et al., 2020, Guan et al., 2021].

References I

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In Proceedings of the 33rd Annual Conference on Learning Theory, volume 125, pages 64–66, 2020.

A. Beck. First-Order Methods in Optimization. SIAM, 2017.

Q. Cai, M. Hong, Y. Chen, and Z. Wang. On the global convergence of imitation learning: A case for linear quadratic regulator. arXiv, 1901.03674, 2019.

R. Foygel and L. W. Mackey. Corrupted sensing: Novel guarantees for separating structured signals. IEEE Transactions on Information Theory, 60(2):1223–1247, 2014.

S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Mathematical Programming, 156(1-2):59–99, 2016.

References II

Z. Guan, T. Xu, and Y. Liang. When will generative adversarial imitation learning algorithms attain global convergence. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, pages 1117–1125, 2021.

S. M. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In Proceedings of the 17th International Conference on Machine Learning, pages 267–274, 2002.

S. Lojasiewicz. A topological property of real analytic subsets. Coll. du CNRS, Les équations aux dérivées partielles, 117:87–89, 1963.

B. T. Polyak. Gradient methods for minimizing functionals. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 3(4):643–653, 1963.

M. L. Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, 2014.

References III

Y. Zhang, Q. Cai, Z. Yang, and Z. Wang. Generative adversarial imitation learning with neural networks: Global optimality and convergence rate. arXiv, 2003.03709, 2020.