Theory and Methods for Reinforcement Learning

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 6: Policy Gradient 2

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-618 (Spring 2022)

License Information for Theory and Methods for Reinforcement Learning (EE-618)

Recap: Policy-based methods

Policy optimization (episodic reward)

$$\max_{\theta} J(\pi_\theta) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 \sim \mu, \pi_\theta\right] = \mathbb{E}_{s\sim\mu}[V^{\pi_\theta}(s)]$$

Tabular parametrization

- Direct :

$$\pi_\theta(a|s) = \theta_{s,a}, \text{ with } \theta_{s,a} \geq 0, \sum\nolimits_a \theta_{s,a} = 1$$

- Softmax:

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'\in\mathcal{A}} \exp(\theta_{s,a'})}$$

Non-tabular parametrization

- Softmax:

$$\pi_\theta(a|s) = \frac{\exp(f_\theta(s,a))}{\sum_{a'\in\mathcal{A}} \exp(f_\theta(s,a'))}$$

- Gaussian:

$$\pi_\theta(a|s) \sim \mathcal{N}\left(\mu_\theta(s), \sigma_\theta^2(s)\right)$$

Recap: Policy gradient theorems

∘ Recall that $\mathrm{p}_\theta(\tau)$ is the trajectory distribution and $\lambda_\mu^\pi(s)$ is the discounted state visitation distribution.

Policy gradient theorems

- REINFORCE expression is given by

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim \mathrm{p}_\theta}\left[R(\tau)\left(\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t)\right)\right].$$

- Action-value expression is given by

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\tau\sim \mathrm{p}_\theta}\left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim\lambda_\mu^{\pi_\theta},a\sim\pi_\theta(\cdot|s)}\left[Q^{\pi_\theta}(s,a)\nabla_\theta \log \pi_\theta(a|s)\right].\end{aligned}$$

Policy gradient in tabular setting

- Direct parametrization: $\pi_\theta(a|s) = \theta_{s,a}$

$$\frac{\partial J(\pi_\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma}\lambda_\mu^{\pi_\theta}(s) Q^{\pi_\theta}(s,a)$$

- Softmax parametrization: $\pi_\theta(a|s) \propto \exp(\theta_{s,a})$

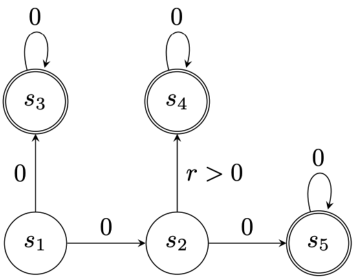
$$\frac{\partial J(\pi_\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma}\lambda_\mu^{\pi_\theta}(s)\pi_\theta(a|s) A^{\pi_\theta}(s,a)$$

Proofs:

- Recall that $\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma}\sum_s \lambda_\mu^{\pi_\theta}(s)\sum_a Q^{\pi_\theta}(s,a)\nabla_\theta \pi_\theta(a|s)$.
- Direct case: $\frac{\partial \pi_\theta(a|s)}{\partial \theta_{s',a'}} = \mathbf{1}\{s=s', a=a'\}$.
- Softmax case: $\frac{\partial \pi_\theta(a|s)}{\partial \theta_{s',a'}} = \pi_\theta(a|s)\mathbf{1}\{s=s', a=a'\} - \pi_\theta(a|s)\pi_\theta(a'|s)\mathbf{1}\{s=s'\}$.

Optimization challenge I: Nonconcavity

- In general, the objective $J(\pi_\theta)$ is nonconcave.
- This holds even for tabular setting with direct or softmax parametrization.



a_1: move up, a_2: move right

Example (direct parametrization)

$$V^\pi(s_1) = \pi(a_2|s_1)\pi(a_1|s_2)r.$$

- Consider $\pi_{\text{mid}} = \frac{\pi_1+\pi_2}{2}$, where

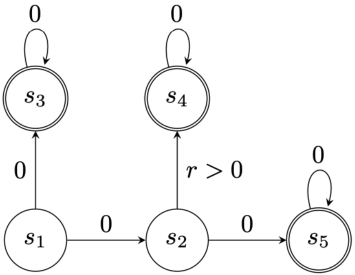
$$\begin{aligned} \pi_1(a_2|s_1) &= 3/4, & \pi_1(a_1|s_2) &= 3/4; \\ \pi_2(a_2|s_1) &= 1/4, & \pi_2(a_1|s_2) &= 1/4; \\ \pi_{\text{mid}}(a_2|s_1) &= 1/2, & \pi_{\text{mid}}(a_1|s_2) &= 1/2. \end{aligned}$$

- $V^{\pi_1}(s_1) = \frac{9}{16}r, V^{\pi_2}(s_1) = \frac{1}{16}r.$
- $V^{\pi_{\text{mid}}}(s_1) = \frac{1}{4}r < \frac{1}{2}(V^{\pi_1}(s_1) + V^{\pi_2}(s_1)).$

Optimization challenge I: Nonconcavity

∘ In general, the objective $J(\pi_\theta)$ is nonconcave.

∘ This holds even for tabular setting with direct or softmax parametrization.



a_1: move up, a_2: move right

Example (softmax parameterzation)

$$\theta = (\theta_{a_1,s_1}, \theta_{a_2,s_1}, \theta_{a_1,s_2}, \theta_{a_2,s_2}),$$

$$V^{\pi_\theta}(s_1) = \frac{e^{\theta_{a_2,s_1}}}{e^{\theta_{a_1,s_1}} + e^{\theta_{a_2,s_1}}} \frac{e^{\theta_{a_1,s_2}}}{e^{\theta_{a_1,s_2}} + e^{\theta_{a_2,s_2}}} r.$$

- Consider

$$\begin{aligned}
\theta_1 &= (\log 1, \log 3, \log 3, \log 1),\\
\theta_2 &= (-\log 1, -\log 3, -\log 3, -\log 1),\\
\theta_{\text{mid}} &= (\theta_1 + \theta_2)/2 = (0, 0, 0, 0).
\end{aligned}$$

- $V^{\pi_{\theta_1}}(s_1) = \frac{9}{16} r, V^{\pi_{\theta_2}}(s_1) = \frac{1}{16} r.$
- $V^{\pi_{\theta_{\text{mid}}}}(s_1) = \frac{1}{4} r < \frac{1}{2}(V^{\pi_{\theta_1}}(s_1) + V^{\pi_{\theta_2}}(s_1)).$

Convergence to stationary points (see Lecture 1)

Convergence of exact policy gradient method: $\theta_{t+1} = \theta_t + \alpha_t \nabla_\theta J(\pi_{\theta_t})$ (Nesterov, 2004 [7])

If the objective $J(\pi_\theta)$ is L-smooth and set $\alpha_t = \frac{1}{L}$, then we have the following guarantee:

$$\min_{t=0,\ldots,T-1} \|\nabla_\theta J(\pi_{\theta_t})\|_2^2 \le \frac{2L(J(\pi_{\theta^\star}) - J(\pi_{\theta_0}))}{T}.$$

Convergence of stochastic policy gradient method: $\theta_{t+1} = \theta_t + \alpha_t \hat{\nabla}_\theta J(\pi_{\theta_t})$ (Ghadimi and Lan, 2013 [3])

If the objective $J(\pi_\theta)$ is L-smooth and $\hat{\nabla}_\theta J(\pi_\theta)$ is unbiased and has bounded variance by σ^2, then with a proper choice of the step-size, we have the following guarantee:

$$\min_{t=0,\ldots,T-1} \mathbb{E}\left[\|\nabla_\theta J(\pi_{\theta_t})\|_2^2\right] = O\left(\sqrt{\frac{L(J(\pi_{\theta^\star}) - J(\pi_{\theta_0}))\sigma^2}{T}}\right).$$

Questions: Can these rates be further improved? Do stationary points imply good performance?

Optimization challenge II: Vanishing gradient and saddle points

- In general, there are no guarantees on the quality of stationary points.
- Vanishing gradients can happen when using softmax parametrization.
- Vanishing gradients can happen when lacking sufficient exploration [1].

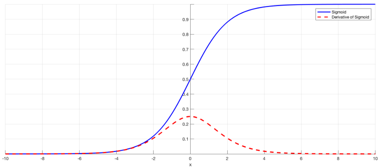


Figure: Softmax function: $\frac{e^\theta}{1+e^\theta} = \frac{1}{1+e^{-\theta}}$.

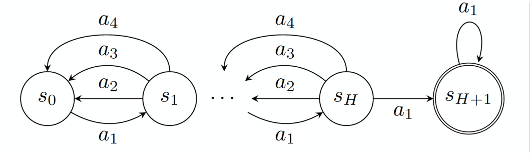


Figure: Example with $H+2$ states and $\gamma = \frac{H}{H+1}$: rewards are everywhere 0 except at s_{H+1}. For small order p and θ such that $\theta_{s,a_1} < \frac{1}{4}$ for all s [1]: $\|\nabla^p V^{\pi_\theta}(s_0)\| \leq \left(\frac{1}{3}\right)^{H/4}$.

A simple example

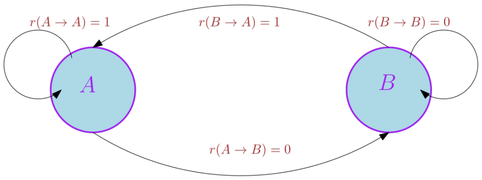


Figure: MDP with 2 states and 2 actions

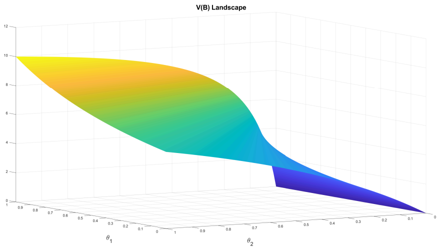


Figure: $V^{\pi}(B)$ under direct parametrization

A simple example (cont'd)

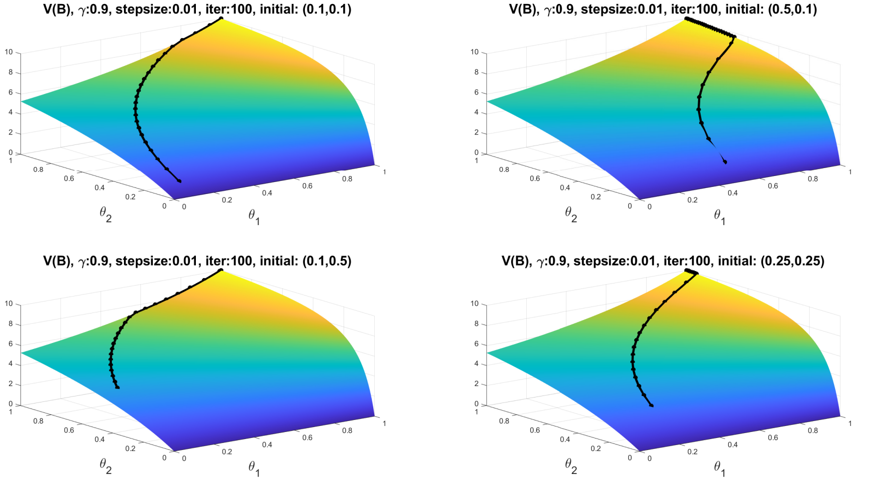


Figure: PG with different initial points

A simple example (cont'd)

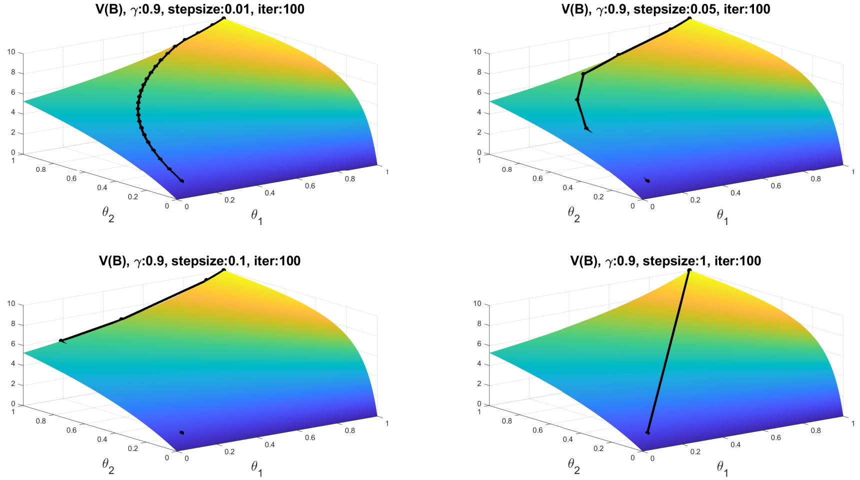


Figure: PG with different stepsizes

Fundamental questions

Question 1

When do policy gradient methods converge to an optimal solution? If so, how fast?

Remarks: ◦ Optimization wisdom: GD/SGD could converge to the global optima for "convex-like" functions:

$$J(\pi^\star) - J(\pi) = O(\|\nabla J(\pi)\|).$$

◦ Focus on tabular setting with exact gradient.

Question 2

How to avoid vanishing gradients and improve the convergence?

Remarks: ◦ Optimization wisdom: Use divergence with good curvature information.

◦ Switch to natural policy gradient by exploiting geometry.

Performance difference lemma (PDL)

Performance difference lemma (Kakade and Langford, 2002 [?])

For any two policy π, π', the following holds

$$J(\pi) - J(\pi') = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\lambda_\mu^\pi,\ a\sim\pi(\cdot|s)}\left[A^{\pi'}(s,a)\right].$$

Remarks:

- Here $\lambda_\mu^\pi(s) = (1-\gamma)\mathbb{E}[\sum_{t=0}^\infty \gamma^t \mathbf{1}_{\{s_t=s\}}|s_0\sim\mu,\pi]$ is the state visitation distribution.
- Here $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ is the advantage function.
- Can be used to show policy improvement theorem for policy iteration (self-exercise).
- Can also be used to show policy gradient theorem (self-exercise).
- Proof follows from definition of value functions.

Proof of performance difference lemma

Derivation:

$$V^{\pi}(s) - V^{\pi'}(s) = \mathbb{E}_{\tau \sim \mathsf{p}_\pi(\tau)}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s\Big] - V^{\pi'}(s)$$

$$= \mathbb{E}_{\tau \sim \mathsf{p}_\pi(\tau)}\Big[\sum_{t=0}^{\infty} \gamma^t \Big(r(s_t, a_t) + V^{\pi'}(s_t) - V^{\pi'}(s_t)\Big) | s_0 = s\Big] - V^{\pi'}(s)$$

$$= \mathbb{E}_{\tau \sim \mathsf{p}_\pi(\tau)}\Big[\sum_{t=0}^{\infty} \gamma^t \Big(r(s_t, a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t)\Big) | s_0 = s\Big]$$

$$= \mathbb{E}_{\tau \sim \mathsf{p}_\pi(\tau)}\Big[\sum_{t=0}^{\infty} \gamma^t \Big(r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim P(\cdot | s_t, a_t)}[V^{\pi'}(s_{t+1})] - V^{\pi'}(s_t)\Big) | s_0 = s\Big]$$

$$= \mathbb{E}_{\tau \sim \mathsf{p}_\pi(\tau)}\Big[\sum_{t=0}^{\infty} \gamma^t \Big(Q^{\pi'}(s_t, a_t) - V^{\pi'}(s_t)\Big) | s_0 = s\Big]$$

$$= \mathbb{E}_{\tau \sim \mathsf{p}_\pi(\tau)}\Big[\sum_{t=0}^{\infty} \gamma^t A^{\pi'}(s_t, a_t) | s_0 = s\Big]$$

Remark: ○ We use a telescoping trick to go from line 2 to line 3!

Key insight: Policy optimization is convex-like in the full policy space

- Performance difference lemma:

$$J(\pi^\star) - J(\pi) = \frac{1}{1-\gamma} \sum_s \lambda_\mu^{\pi^\star}(s) \sum_a \pi^\star(a|s) A^\pi(s,a).$$

- Policy gradient theorem (tabular setting):

$$\frac{\partial J(\pi)}{\partial \pi(a|s)} = \frac{1}{1-\gamma} \lambda_\mu^\pi(s) Q^\pi(s,a) \qquad \text{(direct parametrization).}$$

$$\frac{\partial J(\pi)}{\partial \pi(a|s)} = \frac{1}{1-\gamma} \lambda_\mu^\pi(s) \pi(a|s) A^\pi(s,a) \qquad \text{(softmax parametrization).}$$

- This seems to imply gradient dominance type properties:

$$J(\pi^\star) - J(\pi) = O(\|\nabla J(\pi)\|),$$

which is crucial to ensure global optimality.

Policy optimization

◦ We first consider the direct parametrization in the tabular setting.

Policy optimization under direct parametrization

$$\max_{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|}} J(\pi) := \mathbb{E}_{s\sim\mu}[V^\pi(s)],$$

where $\Delta(\mathcal{A})^{|\mathcal{S}|} = \{\pi : \pi(a|s) \geq 0, \sum_{a\in\mathcal{A}} \pi(a|s) = 1, \forall s\}$. For brevity, we denote this set as Δ.

Remarks:

◦ If $\pi \in \Delta$ is optimal, then it satisfies the first-order optimality condition:

$$\langle \bar{\pi} - \pi, \nabla J(\pi)\rangle \leq 0, \forall\, \bar{\pi} \in \Delta,$$

or equivalently, $\max_{\bar{\pi}\in\Delta} \langle \bar{\pi} - \pi, \nabla J(\pi)\rangle = 0$.

◦ Does the reverse statement hold?

Gradient dominance property

Gradient mapping domination

$$J(\pi^\star) - J(\pi) \le \left\| \frac{\lambda_\mu^{\pi^\star}}{\lambda_\mu^\pi} \right\|_\infty \times \max_{\bar{\pi}\in\Delta} \langle \bar{\pi} - \pi, \nabla J(\pi) \rangle.$$

Remarks:

- Any first-order stationary point is thus globally optimal.
- The term $\left\| \frac{\lambda_\mu^{\pi^\star}}{\lambda_\mu^\pi} \right\|_\infty$ is called the distribution mismatch coefficient, which captures the hardness of the exploration problem. Note that in the aforementioned vanishing gradient example, this coefficient can be very exponentially large.
- Note that $\max_\pi \left\| \frac{\lambda_\mu^{\pi^\star}}{\lambda_\mu^\pi} \right\|_\infty \le \frac{1}{1-\gamma} \left\| \frac{\lambda_\mu^{\pi^\star}}{\mu} \right\|_\infty$, since $\forall \pi, \lambda_\mu^\pi(s) \ge (1-\gamma)\mu(s)$.
- Proof follows by combining performance difference lemma and policy gradient theorem.

Proof of gradient dominance

Derivation:

$$
\begin{aligned}
J(\pi^\star) - J(\pi) &= \frac{1}{1-\gamma} \sum_s \lambda_\mu^{\pi^\star}(s) \sum_a \pi^\star(a|s) A^\pi(s,a) \\
&= \frac{1}{1-\gamma} \sum_s \frac{\lambda_\mu^{\pi^\star}(s)}{\lambda_\mu^\pi(s)} \lambda_\mu^\pi(s) \sum_a \pi^\star(a|s) A^\pi(s,a) \\
&\leq \frac{1}{1-\gamma} \left\| \frac{\lambda_\mu^{\pi^\star}}{\lambda_\mu^\pi} \right\|_\infty \times \max_{\bar{\pi}\in\Delta} \sum_{s,a} \lambda_\mu^\pi(s) \bar{\pi}(a|s) A^\pi(s,a) \\
&= \frac{1}{1-\gamma} \left\| \frac{\lambda_\mu^{\pi^\star}}{\lambda_\mu^\pi} \right\|_\infty \times \max_{\bar{\pi}\in\Delta} \sum_{s,a} \lambda_\mu^\pi(s) (\bar{\pi}(a|s) - \pi(a|s)) A^\pi(s,a) \\
&\leq \frac{1}{1-\gamma} \left\| \frac{\lambda_\mu^{\pi^\star}}{\lambda_\mu^\pi} \right\|_\infty \times \max_{\bar{\pi}\in\Delta} \sum_{s,a} \lambda_\mu^\pi(s) (\bar{\pi}(a|s) - \pi(a|s)) Q^\pi(s,a) \\
&= \left\| \frac{\lambda_\mu^{\pi^\star}}{\lambda_\mu^\pi} \right\|_\infty \times \max_{\bar{\pi}\in\Delta} \langle \bar{\pi} - \pi, \nabla J(\pi) \rangle
\end{aligned}
$$

Projected policy gradient method

Projected policy gradient method

$$\pi_{t+1} = \Pi_{\Delta}(\pi_t + \eta \nabla J(\pi_t)),$$

where the projection is given by $\Pi_{\Delta}(\pi) = \arg\min_{\pi' \in \Delta} \|\pi - \pi'\|_2^2$.

Remarks:

- Take a gradient ascent step and project onto the simplex set (can be computed efficiently).
- *Generalized gradient mapping*: $G(\pi_t) = \frac{1}{\eta}(\pi_{t+1} - \pi_t)$, or equivalently, $\pi_{t+1} = \pi_t + \eta G(\pi_t)$.
- If π is optimal, then $G(\pi) = 0$. (why?)
- Convergence on gradient mapping [6]: If $J(\pi)$ is L-smooth, then we have

$$\min_{t \le T} \|G(\pi_t)\|_2^2 \le \frac{2L(J(\pi^\star) - J(\pi_0))}{T}.$$

Convergence of projected policy gradient method

Theorem (Agarwal et al., 2020 [1])

Assume access to exact gradient. Let $\eta = \frac{(1-\gamma)^3}{2\gamma|\mathcal{A}|}$. Then, the following holds

$$\min_{t<T} J(\pi^\star) - J(\pi_t) \leq \frac{8\sqrt{\gamma|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^3\sqrt{T}} \left\| \frac{\lambda_\mu^{\pi^\star}}{\mu} \right\|_\infty .$$

Proof sketch:

- Show that the objective $J(\pi)$ is L-smooth with $L = \frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3}$ and $J(\pi) \leq \frac{1}{1-\gamma}$.
- Invoke convergence on gradient mapping: $\min_{t\leq T} \|G(\pi_t)\|_2^2 \leq \frac{2L(J(\pi^\star)-J(\pi_0))}{T}$.
- Invoke the relationship between gradient mapping and approximation of stationary point [6]:

$$\max_{\bar{\pi}\in\Delta} \langle \bar{\pi} - \pi_{t+1}, \nabla J(\pi_{t+1}) \rangle \leq (1 + L\eta) \cdot \|G(\pi_t)\|_2 \cdot \|\pi_{t+1} - \pi_t\|_2.$$

- Use the gradient dominance for global convergence.

A closer look at the convergence

Theorem (Agarwal et al., 2020 [1])

Assume access to exact gradient. Let $\eta = \frac{(1-\gamma)^3}{2\gamma|A|}$. Then, the following holds

$$\min_{t<T} J(\pi^\star) - J(\pi_t) \leq \frac{8\sqrt{\gamma|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^3\sqrt{T}} \left\| \frac{\lambda_\mu^{\pi^\star}}{\mu} \right\|_\infty .$$

Remarks:

- Large constants in the bound.
- Slow rate in T.
- Analysis can be refined with improved convergence rate of $O\left(\frac{1}{T}\right)$ using Nesterov's result in (Nesterov, 2004 [7]).
- But wait, in tabular setting, VI or PI converges linearly, which is much faster.
- (New!) Linear convergence of PG can be shown with larger stepsizes (through line-search) (Bhandari and Russo, 2021 [2]).

A closer look at the PG method

- The projected PG update can also be viewed as

$$\pi_{t+1} := \Pi_\Delta(\pi_t + \eta \nabla J(\pi_t)) \\ = \arg\max_{\pi \in \Delta} \left\{ \langle \nabla J(\pi_t), \pi \rangle - \frac{1}{2\eta} \|\pi - \pi_t\|_2^2 \right\}.$$

- As $\eta \to \infty$, this reduces to the policy iteration update:

$$\pi_{t+1}(\cdot|s) = \arg\max_{\pi(\cdot|s) \in \Delta(\mathcal{A})} \sum_a \pi(s|a) Q^{\pi_t}(s,a).$$

- In other words, policy gradient method can be viewed as an approximation of policy iteration

$$\arg\max_{\pi \in \Delta} \left\{ \langle \nabla J(\pi_t), \pi \rangle - \frac{1}{2\eta} \|\pi - \pi_t\|_2^2 \right\} = \arg\max_{\pi \in \Delta} \left\{ \langle Q^{\pi_t}, \pi \rangle_{\lambda_\mu^{\pi_t}} - \frac{1}{2\eta'} \|\pi - \pi_t\|_2^2 \right\} \quad (1)$$

where $\frac{\partial J(\pi)}{\partial \pi(a|s)} = \frac{1}{1-\gamma} \lambda_\mu^\pi(s) Q^\pi(s,a)$ and $\langle \cdot, \cdot \rangle_{\lambda_\mu^\pi}$ is the reweighted inner product by λ_μ^π.

From gradient descent to mirror descent: Exploiting the non-euclidean geometry

- We can adapt PG in the simplex with mirror descent updates:

$$\pi_{t+1} := \arg\max_{\pi \in \Delta} \left\{ \langle \nabla J(\pi_t), \pi \rangle - \frac{1}{\eta} \sum_s \lambda_\mu^{\pi_t}(s) \mathsf{KL}\left(\pi(\cdot|s) || \pi_t(\cdot|s)\right) \right\},$$

where $\mathsf{KL}(p||q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$ is the Kullback-Leibler divergence.

- The policy mirror descent update can be further simplified as

$$\pi_{t+1}(a|s) = \pi_t(a|s) \frac{\exp(\eta Q^t(s,a)/(1-\gamma))}{\sum_{a'} \pi_t(a'|s) \exp(\eta Q^t(s,a')/(1-\gamma))}.$$

- This is akin to natural policy gradient under softmax parameterization.

- As $\eta \to \infty$, this also reduces to the policy iteration update.

Policy optimization

◦ We now consider the softmax parametrization in the tabular setting.

Policy optimization under softmax parametrization

$$\max_{\theta} J(\pi_\theta) := \mathbb{E}_{s\sim\mu}[V^{\pi_\theta}(s)], \quad \text{where } \pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}.$$

Softmax policy gradient method

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta J(\pi_{\theta_t}), \quad \text{where} \quad \frac{\partial J(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} \lambda_\mu^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi_\theta}(s,a).$$

Gradient dominance and global convergence

Gradient dominance (Mei et al., 2020 [5])

$$J(\pi^\star) - J(\pi_\theta) \leq [\min_s \pi_\theta(a^\star(s)|s)]^{-1} \sqrt{S} \cdot \left\| \frac{\lambda_\mu^{\pi^\star}}{\lambda_\mu^{\pi_\theta}} \right\|_\infty \cdot \|\nabla_\theta J(\pi_\theta)\|_2.$$

Convergence of softmax policy gradient (Mei et al., 2020 [5])

Assume access to exact gradient, let $\eta \leq \frac{(1-\gamma)^3}{8}$. Then, the following holds

$$J(\pi^\star) - J(\pi_{\theta_T}) \leq \frac{16|\mathcal{S}|}{c^2(1-\gamma)^5 T} \left\| \frac{\lambda_\mu^{\pi^\star}}{\mu} \right\|_\infty^2,$$

where $c = [\min_{s,t} \pi_{\theta_t}(a^\star(s)|s)]^{-1} > 0$.

Remark: ◦ Proof follows similarly as the tabular setting with slow rate and large constants in the bound.

Natural policy gradient method (NPG)

Natural policy gradient (Kakade, 2002 [4])

$$\theta_{t+1} = \theta_t + \eta (F_{\theta_t})^{\dagger} \nabla J(\pi_{\theta_t}),$$

where

- F_θ is the Fisher information matrix:

$$F_\theta = \mathbb{E}_{s \sim \lambda_\mu^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^\top \right].$$

- $C^\dagger$ is the pseudoinverse of the matrix C.

NPG under softmax parameterization

- Consider $\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$ and denote $\pi_t = \pi_{\theta_t}$.

NPG parameter update

$$\theta_{t+1} = \theta_t + \frac{\eta}{1-\gamma} A^{\pi_{\theta_t}}.$$

NPG policy update = policy mirror descent

$$\pi_{t+1}(a|s) = \pi_t(a|s) \frac{\exp(\eta A^{\pi_t}(s,a)/(1-\gamma))}{\sum_{a'} \pi_t(a'|s) \exp(\eta A^{\pi_t}(s,a')/(1-\gamma))}.$$

Convergence of NPG

Convergence of NPG with softmax parameterization [1]

Assume access to A^{π_θ}. For any $\eta \geq (1-\gamma)^2 \log|\mathcal{A}|$ and $T > 0$, we have the following

$$J(\pi^\star) - J(\pi_{\theta_T}) \leq \frac{2}{(1-\gamma)^2 T}.$$

Remarks:

- ◦ Dimension-free convergence, no dependence on $|\mathcal{A}|, |\mathcal{S}|$.
- ◦ No dependence on distribution mismatch coefficient.

Questions: Why? What about function approximation setting? Can we further improve the convergence?

Next week!

- Recap on policy gradient methods
- Introduction to natural policy gradient method

References I

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan.
Optimality and approximation with policy gradient methods in markov decision processes.
In *Conference on Learning Theory*, pages 64–66. PMLR, 2020.

[2] Jalaj Bhandari and Daniel Russo.
On the linear convergence of policy gradient methods for finite mdps.
In *International Conference on Artificial Intelligence and Statistics*, pages 2386–2394. PMLR, 2021.

[3] Saeed Ghadimi and Guanghui Lan.
Stochastic first- and zeroth-order methods for nonconvex stochastic programming.
SIAM Journal on Optimization, 23(4):2341–2368, 2013.

[4] S. Kakade.
A natural policy gradient.
In *Advances in Neural Information Processing Systems (NeurIPS)*, 2001.

[5] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans.
On the global convergence rates of softmax policy gradient methods.
In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

[6] Yu Nesterov.
Gradient methods for minimizing composite functions.
Mathematical Programming, 140(1):125–161, 2013.

[7] Yurii Nesterov.
Introductory Lectures on Convex Optimization.
Kluwer, Boston, MA, 2004.