On the Linear Convergence rate of Policy Gradient methods

Tian Xu

School of Artificial Intelligence
Nanjing University

September 24, 2020

RL theory reading group
Mainly based on paper:
https://arxiv.org/abs/2007.11120

Markov Decision Process

- Consider an infinite-horizon discounted Markov Decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, g, P, \gamma, \rho)$.
  - $\mathcal{S}$ and $\mathcal{A}$ are the state and action space, respectively.
  - g denotes the cost function.
  - P specifies the transition probability of s_{t+1} conditioned on s_t and a_t.
  - $\gamma \in [0, 1)$ is a discount factor.
  - ρ determines the initial state distribution.

Markov Decision Process

- We focus on finite state space $\mathcal{S} = \{s_1, s_2, \cdots, s_n\}$. For each state $s_i \in \mathcal{S}$, there is a finite set of k actions to choose from.
- Let $\mathcal{A} = \Delta^{k-1}$ be the set of all probability distributions over k actions and $a \in \mathcal{A}$ is a probability vector where each component a_i denotes the probability of taking $i^{\textbf{th}}$ action.
- A stationary policy $\pi : \mathcal{S} \to \mathcal{A}$ and we use $\pi(s, i)$ denote the $i^{\textbf{th}}$ component of $\pi(s)$. $\Pi = \mathcal{A}^n$ denotes the set of all stationary policies.

Markov Decision Process

- Given policy $\pi \in \Pi$, the cost to go function $J_\pi : \mathcal{S} \to \mathbb{R}$ is defined as

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

- Given policy $\pi \in \Pi$, the Bellman operator $T_\pi : \mathbb{R}^n \to \mathbb{R}^n$ is defined as:

$$(T_\pi J)(s) := g(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) J(s')$$

- The cost to go function of policy π is the unique fixed point of T_π:

$$J_\pi = T_\pi J_\pi$$

Markov Decision Process

- The Bellman optimality operator $T : \mathbb{R}^n \to \mathbb{R}^n$ is defined as

$$(TJ)(s) := \min_{\pi \in \Pi} g(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s))J(s') = \min_{\pi \in \Pi} (T_\pi J)(s)$$

- The optimal cost-to-go function $J^*(s) = \min_\pi J_\pi(s)$ is the unique function of T:

$$J^* = TJ^*$$

Markov Decision Process

- The state-action cost-to-go function of a policy $\pi \in \Pi$:

$$Q_\pi(s,a) = g(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) J_\pi(s').$$

- The relationship between Q_π, J_π, T_π and T:

$$Q_\pi(s,\pi(s)) = J_\pi(s) \quad Q_\pi(s,\pi'(s)) = (T_{\pi'} J_\pi)(s) \quad \min_{a \in \mathcal{A}} Q_\pi(s,a) = (TJ_\pi)(s)$$

Markov Decision Process

- The loss function of policy gradient methods:

$$l(\pi) = (1-\gamma)\sum_{s\in\mathcal{S}} J_\pi(s)\rho(s),$$

where ρ is the initial state distribution.

- Under the assumption that $\rho(s) > 0 \quad \forall s \in \mathcal{S}$,

$$\pi \in \operatorname{argmin}_{\bar{\pi}} l(\bar{\pi}) \iff \pi \in \operatorname{argmin}_{\bar{\pi}} J_{\bar{\pi}}(s) \quad \forall s \in \mathcal{S}.$$

- The discounted state-occupancy measure under π and ρ is defined as

$$\eta_\pi(\cdot) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = \cdot) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \rho P_\pi^t = (1-\gamma)\rho(I-\gamma P_\pi)^{-1}.$$

where $P_\pi = (P(s'|s,\pi(s)))_{s,s'\in\mathcal{S}} \in \mathbb{R}^{n\times n}$.

Policy Iteration

- Starting with policy π, policy iteration (PI) performs the following steps iteratively:
- Policy Evaluation: calculate Q_π by performing Bellman operator T_π.
- Policy Improvement: find the greedy policy π^+ corresponding to Q_π:

$$\pi^+(s) \in \mathrm{argmin}_{a \in \mathcal{A}} Q_\pi(s, a).$$

- PI enjoys the linear convergence rate:

$$\|J_{\pi^+} - J_*\|_\infty \leq \gamma \|J_\pi - J_*\|_\infty$$

Proof of the linear convergence rate

- Given a policy $\pi \in \Pi$, the Bellman operator T_π and the Bellman optimality operator T have the following properties:
- Monotonicity: $\forall J_1, J_2 \in \mathbb{R}^n \quad \text{s.t. } J_1 \le J_2$, then it holds that $T_\pi J_1 \le T_\pi J_2, \ TJ_1 \le TJ_2$.
- γ-contraction: $\forall J_1, J_2 \in \mathbb{R}^n$, it holds that $\|T_\pi J_1 - T_\pi J_2\|_\infty \le \gamma \|J_1 - J_2\|_\infty, \quad \|TJ_1 - TJ_2\|_\infty \le \gamma \|J_1 - J_2\|_\infty$.

Proof of the linear convergence rate

- Starting with policy π, π^+ acts greedily with respect to $Q_\pi(s, a)$.
- $T_{\pi^+} J_\pi = T J_\pi = \min_{\bar{\pi} \in \Pi} T_{\bar{\pi}} J_\pi \leq T_\pi J_\pi = J_\pi$.
- $J_\pi \geq T_{\pi^+} J_\pi \geq T^2_{\pi^+} J_\pi \geq \cdots \geq J_{\pi^+}$.
- $\| J_{\pi^+} - J^* \|_\infty = \| T_{\pi^+} J_{\pi^+} - J^* \|_\infty \leq \| T_{\pi^+} J_\pi - J^* \|_\infty \leq \| T J_\pi - T J^* \|_\infty \leq \gamma \| J_\pi - J^* \|_\infty$

Policy space v.s. Paramterization space

- In policy gradient methods, we often parametrize policy π_θ with θ and consider the gradient of $I(\pi_\theta)$ with respect to θ.
  - For example, we consider a softmax policy π_θ defined by $\pi_\theta(s,i) \propto \exp(\theta_{s,i})$, where $\theta \in \mathbb{R}^{n\times k}$.
- In this talk, we focus on policy gradients directly on the policy space $(\pi(s,i))_{n\times k}$.
  - When we use direct policy parameterization that $\pi_\theta(s,i) = \theta_{s,i}$ s.t. $\sum_{i\in[k]} \theta_{s,i} = 1\ \forall s \in \mathcal{S}$, the policy gradient w.p.t $\pi(s,i)$ is equivalent to the policy gradient w.p.t. $\theta_{s,i}$.
  - Mathematical analysis is much cleaner over the policy space since it is closed.

Connection between policy gradient and policy iteration

- Define the weighted policy iteration objective:

$$\mathcal{B}(\bar{\pi}|\eta, J_\pi) = \sum_{s=1}^{n} \eta(s) \sum_{i=1}^{k} Q_\pi(s,i)\bar{\pi}(s,i) = \sum_{s=1}^{n} \eta(s)(T_{\bar{\pi}} J_\pi)(s) = \langle Q_\pi, \bar{\pi} \rangle_{\eta \times 1}$$

  where $\langle v, u \rangle_W = \sum_{s=1}^{n} \sum_{i=1}^{k} v(s,i)u(s,i)W(s,i)$.
- If the state distribution η supports on the entire state space, then we have

$$\pi^+ \in \mathrm{argmin}_{\bar{\pi} \in \Pi} \mathcal{B}(\bar{\pi}|\eta, J_\pi) \iff \pi^+(s) \in \mathrm{argmin}_{a \in \mathcal{A}} Q_\pi(s,a).$$

Connection between policy gradient and policy iteration

- The gradients of the cost function $I(\pi) = \sum_{s \in \mathcal{S}} \rho(s) J_\pi(s)$ equal the gradients of the weighted policy iteration objective $\mathcal{B}(\bar{\pi} | \eta_\pi, J_\pi) = \sum_{s=1}^{n} \eta_\pi(s) \sum_{i=1}^{k} Q_\pi(s,i)\bar{\pi}(s,i)$:

$$
\begin{aligned}
\nabla_\pi I(\pi) &= \mathbb{E}_{s \sim \eta_\pi(\cdot), i \sim \pi(\cdot|s)}[\nabla_\pi \log \pi(s,i) Q_\pi(s,i)] \\
&= \sum_{s,i} \eta_\pi(s)\pi(s,i) \begin{bmatrix} 0 \\ \vdots \\ \frac{1}{\pi(s,i)} \\ \vdots \\ 0 \end{bmatrix} Q_\pi(s,i) \\
&= (\eta_\pi(s) Q_\pi(s,i))_{s \in \mathcal{S}, i \in [k]} \\
&= \nabla_{\bar{\pi}} \mathcal{B}(\bar{\pi} | \eta_\pi, J_\pi)
\end{aligned}
$$

Frank-Wolfe Algorithm

- Starting with policy $\pi \in \Pi$, an iteration of the Frank-Wolfe method performs the following two steps:
- Linear optimization:

$$\begin{aligned}\pi^{+} &= \mathrm{argmin}_{\bar{\pi}\in\Pi}\langle \nabla_{\pi} l(\pi), \bar{\pi}\rangle = \mathrm{argmin}_{\bar{\pi}\in\Pi}\sum_{s}\eta_{\pi}(s)\sum_{i=1}^{k} Q_{\pi}(s,i)\bar{\pi}(s,i) \\ &= \mathrm{argmin}_{\bar{\pi}\in\Pi}\mathcal{B}(\bar{\pi}|\eta_{\pi}, J_{\pi})\end{aligned}$$

- Line search and update:

$$\pi' = (1-\alpha)\pi + \alpha\pi^{+} \quad \alpha \in (0,1].$$

- When $\alpha = 1$, the update of Frank-Wolfe method is exactly the update of policy iteration.

Projected Gradient Descent

- Starting with policy $\pi \in \Pi$, the update of projected gradient descent:

$$\begin{aligned} \pi' &= \mathrm{argmin}_{\bar{\pi} \in \Pi} \|\bar{\pi} - (\pi - \alpha \nabla_\pi I(\pi))\|_2^2 \\ &= \mathrm{argmin}_{\bar{\pi} \in \Pi} \langle \nabla_\pi I(\pi), \bar{\pi} \rangle + \frac{1}{2\alpha} \|\bar{\pi} - \pi\|_2^2 \\ &= \mathrm{argmin}_{\bar{\pi} \in \Pi} \mathcal{B}(\bar{\pi} | \eta_\pi, J_\pi) + \frac{1}{2\alpha} \|\bar{\pi} - \pi\|_2^2 \end{aligned}$$

- π' converges to policy iteration when $\alpha \to \infty$.

Mirror-descent

- Instead of using the squared Euclidean penalty $\frac{1}{2\alpha}\|\bar{\pi}-\pi\|_2^2$, Mirror-descent method uses the KL divergence $D_{\mathrm{KL}}(\bar{\pi}(s)||\pi(s))$:

$$\pi' = \mathrm{argmin}_{\bar{\pi}\in\Pi}\langle\nabla J(\pi),\bar{\pi}\rangle + \frac{1}{\alpha}\sum_{s=1}^{n} D_{\mathrm{KL}}(\bar{\pi}(s)||\pi(s))$$

$$= \mathrm{argmin}_{\bar{\pi}\in\Pi}\mathcal{B}(\bar{\pi}|\eta_\pi, J_\pi) + \frac{1}{\alpha}\sum_{s=1}^{n} D_{\mathrm{KL}}(\bar{\pi}(s)||\pi(s))$$

- The closed-form solution:

$$\pi'(s,i) = \frac{\pi(s,i)\exp\{-\alpha\eta_\pi(s)Q_\pi(s,i)\}}{\sum_{j=1}^{k}\pi(s,j)\exp\{-\alpha\eta_\pi(s)Q_\pi(s,j)\}}$$

$$= \frac{\pi(s,i)}{\sum_{j=1}^{k}\pi(s,j)\exp\{\alpha\eta_\pi(s)(Q_\pi(s,i)-Q_\pi(s,j))\}}$$

- When $\alpha\to\infty$, $\pi'(s,i) = \mathrm{argmin}_i Q_\pi(s,i) \quad \forall s\in\mathcal{S}$.

Natural Policy Gradient

- Starting with policy $\pi \in \Pi$, natural policy gradient method penalizes changes to the action distribution at states in proportion to η_π:

$$\pi' = \mathrm{argmin}_{\bar{\pi} \in \Pi} \langle \nabla l(\pi), \bar{\pi} \rangle + \frac{1}{\alpha} \sum_{s=1}^{n} \eta_\pi(s) D_{\mathrm{KL}}(\bar{\pi}(s) || \pi(s))$$

$$= \mathrm{argmin}_{\bar{\pi} \in \Pi} \mathcal{B}(\bar{\pi} | \eta_\pi, J_\pi) + \frac{1}{\alpha} \sum_{s=1}^{n} \eta_\pi(s) D_{\mathrm{KL}}(\bar{\pi}(s) || \pi(s))$$

$$= \left(\frac{\pi(s,i) \exp\{-\alpha Q_\pi(s,i)\}}{\sum_{j=1}^{k} \pi(s,j) \exp\{-\alpha Q_\pi(s,j)\}} \right)_{s \in \mathcal{S}, i \in [k]}.$$

- When $\alpha \to \infty$, $\pi'(s,i) = \mathrm{argmin}_i Q_\pi(s,i) \quad \forall s \in \mathcal{S}$.

The Choice of step-size

- We consider an idealized step-size rule using exact line search. In the step t, we calculate
$$\pi^{t+1} = \text{argmin}_{\pi \in \Pi^{t+1}} \ell(\pi)$$
where $\Pi^{t+1} = \text{Closure}(\{\pi_\alpha^{t+1}\})$ denotes the curve of policies traced out by varying α.
- For Frank-Wolfe method, $\Pi^{t+1} = \{(1-\alpha)\pi^t + \alpha\pi_+^t : \alpha \in (0,1]\}$ is the line segment connecting the current policy π^t and its policy iteration update π_+^t.
- For projected gradient descent, mirror-descent and natural policy gradient, $\Pi^{t+1} = \{\pi_\alpha^{t+1}\}$ is a curve where $\pi_0^{t+1} = \pi^t$ and $\pi_\alpha^{t+1} \to \pi_+^t$ as $\alpha \to \infty$.

The Linear Convergence

Theorem

Suppose one of the policy gradient methods above is applied to minimize $l(\pi)$ over $\pi \in \Pi$. Let π^0 denote the initial policy and $(\pi^t : t \in \{0, 1, 2, \cdots\})$ denote the sequence of iterates. The following bounds holds:

- **Exact line search.** *If the step-sizes are chosen by exact line search, then we have*

$$\|J_{\pi^t} - J^*\|_\infty \leq \left(1 - \min_{s \in \mathcal{S}} \rho(s)(1-\gamma)\right)^t \frac{\|J_{\pi^0} - J^*\|_\infty}{\min_{s \in \mathcal{S}} \rho(s)}$$

- **Constant step-size Frank-Wolfe.** *Under Frank-Wolfe with constant step-size* $\alpha \in (0, 1]$,

$$\|J_{\pi^t} - J^*\|_\infty \leq (1 - \alpha(1-\gamma))^t \|J_{\pi^0} - J^*\|_\infty$$

Proof of exact line search case

- Under each algorithm and at each iteration t, the policy iteration update π_+^t is contained in the policy class Π^{t+1}. we have that

$$\ell(\pi^{t+1}) = \min_{\pi \in \Pi^{t+1}} \ell(\pi) \leq \ell(\pi_+^t)$$

- Recall the property that $J^* \leq J_{\pi_+^t} \leq TJ_{\pi^t} \leq J_{\pi^t}$, we have

$$\begin{aligned}
\ell(\pi^t) - \ell(\pi^{t+1}) \geq \ell(\pi^t) - \ell(\pi_+^t) &= \sum_s \rho(s) \left(J_{\pi^t}(s) - J_{\pi_+^t}(s) \right) \\
&\geq \rho_{\min} \| J_{\pi^t} - J_{\pi_+^t} \|_\infty \\
&\geq \rho_{\min} \| J_{\pi^t} - TJ_{\pi^t} \|_\infty \\
&\geq \rho_{\min} \| J_{\pi^t} - J^* - (TJ_{\pi^t} - J^*) \|_\infty \\
&\geq \rho_{\min} (\| J_{\pi^t} - J^* \|_\infty - \| TJ_{\pi^t} - TJ^* \|_\infty) \\
&\geq \rho_{\min} (1 - \gamma) \| J_{\pi^t} - J^* \|_\infty \\
&\geq \rho_{\min} (1 - \gamma) (\ell(\pi^t) - \ell(\pi^*))
\end{aligned}$$

Proof of exact line search case

- Rearranging terms gives

$$
\begin{aligned}
I(\pi^{t+1}) - I(\pi^*) &\le (1 - \rho_{\min}(1-\gamma))\left(I(\pi^t) - I(\pi^*)\right) \le \cdots \\
&\le (1 - \rho_{\min}(1-\gamma))^t \left(I(\pi^0) - I(\pi^*)\right) \le (1 - \rho_{\min}(1-\gamma))^t \|J_{\pi^0} - J^*\|_\infty
\end{aligned}
$$

- The final result follows from that $\|J_{\pi^t} - J^*\|_\infty \le \frac{\left(I(\pi^{t+1}) - I(\pi^*)\right)}{\rho_{\min}}$

Proof of constant stepsize case

- The Frank-Wolfe update exactly equals a soft-policy iteration update:

$$\pi^{t+1}(s) = (1-\alpha)\pi^{t}(s) + \alpha\pi_{+}^{t}(s)$$

  where π_{+}^{t} is the policy iteration update to π^{t}.
- By the linearity of Bellman operator, for any state s,

$$\begin{aligned}(T_{\pi^{t+1}}J_{\pi^t})(s) &= (1-\alpha)(T_{\pi^t}J_{\pi^t})(s) + \alpha(TJ_{\pi^t})(s) \\ &= (1-\alpha)J_{\pi^t}(s) + \alpha(TJ_{\pi^t})(s) \leq J_{\pi^t}(s)\end{aligned}$$

- By the monotonicty of $T_{\pi^{t+1}}$, we have

$$J_{\pi^t} \geq T_{\pi^{t+1}}J_{\pi^t} \geq T_{\pi^{t+1}}^2 J_{\pi^t} \geq \cdots \geq J_{\pi^{t+1}}$$

Proof of constant stepsize case

- From $J_{\pi^{t+1}} \leq J_{\pi^t}$, it holds that

$$J_{\pi^{t+1}} = T_{\pi^{t+1}} J_{\pi^{t+1}} \leq T_{\pi^{t+1}} J_{\pi^t} = (1-\alpha) J_{\pi^t} + \alpha T J_{\pi^t}$$

- Subtracting J^* from both sides shows

$$J_{\pi^{t+1}} - J^* \leq (1-\alpha)(J_{\pi^t} - J^*) + \alpha(TJ_{\pi^t} - J^*)$$

- By the contraction property of T, we have that

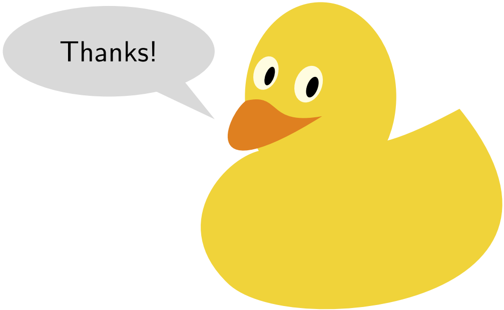
$$\|J_{\pi^{t+1}} - J^*\|_\infty \leq (1 - \alpha + \gamma\alpha)\|J_{\pi^t} - J^*\|_\infty$$

- Applying the inequality obtains the final result:

$$\|J_{\pi^{t+1}} - J^*\|_\infty \leq (1 - \alpha(1-\gamma))^t \|J_{\pi^0} - J^*\|_\infty$$

Questions and Discussions

Any questions or discussions about this talk?


Thanks!