On the Convergence Rates of Policy Gradient Methods

Lin Xiao
Meta AI

40th International Conference on Machine Learning (ICML)

Honolulu, Hawai‘i
July 23-29, 2023

Introduction

- **Markov decision process (MDP)**: $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$
  - $\mathcal{S}$, $\mathcal{A}$: finite state and action spaces
  - $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: transition probability function $P(s'|s, a)$
  - $R : \mathcal{S} \times \mathcal{A} \to \mathbf{R}$: reward (regret) function, with range $[0, 1]$
  - $\gamma \in (0, 1)$: discount factor
- **value (cost) function** with initial state distribution ρ

$$V_\rho(\pi) := \mathop{\mathbf{E}}_{s\sim\rho} V_s(\pi) = \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \,\middle|\, s_0 \sim \rho\right]$$

- **projected policy gradient method** ($\Pi := \Delta(\mathcal{A})^{|\mathcal{S}|}$)

$$\pi^{(k+1)} = \mathbf{proj}_\Pi\left(\pi^{(k)} - \eta_k \nabla V_\mu(\pi^{(k)})\right)$$

- **policy mirror-descent (PMD) method**

$$\pi^{(k+1)} = \mathop{\arg\min}_{\pi\in\Pi}\left\{\eta_k \left\langle \nabla V_\mu(\pi^{(k)}),\, \pi\right\rangle + D_k(\pi, \pi^{(k)})\right\}$$

MDP fundamentals

- **state-visitation distribution**

$$d_{s,s'}(\pi) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr^{\pi}\left(s_t = s' \,|\, s_0 = s\right)$$

- **Q-function** (state-action value function)

$$Q_{s,a}(\pi) := \mathbf{E}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t, a_t) \,\middle|\, s_0 = s,\ a_0 = a\right]$$

- **policy gradient** (Sutton, McAllester, Singh, Mansour 2000)

$$\nabla_s V_\rho(\pi) = \frac{\partial V_\rho(\pi)}{\partial \pi_s} = \frac{1}{1-\gamma} d_{\rho,s}(\pi) Q_s(\pi)$$

- **performance difference lemma** (Kakade & Langford 2002)

  for any $\pi, \tilde{\pi} \in \Pi$ and $s \in \mathcal{S}$

$$V_s(\pi) - V_s(\tilde{\pi}) = \frac{1}{1-\gamma}\,\mathbf{E}_{s'\sim d_s(\pi)}\left\langle Q_{s'}(\tilde{\pi}), \pi_{s'} - \tilde{\pi}_{s'}\right\rangle$$

Projected policy gradient method

- **decoupled form**

$$\pi_s^{(k+1)} = \mathbf{proj}_{\Delta(\mathcal{A})}\left(\pi_s^{(k)} - \eta_k \nabla_s V_\mu(\pi^{(k)})\right), \qquad s \in \mathcal{S}$$

- **key properties** (Agarwal, Kakade, Lee, Mahajan 2021)
  - **smoothness**

$$\|\nabla V_\rho(\pi) - \nabla V_\rho(\pi')\|_2 \leq \frac{2\gamma|\mathcal{A}|}{(1-\gamma)^3}\|\pi - \pi'\|_2$$

  - **variational gradient domination**

$$V_\rho(\pi) - V_\rho^\star \leq \frac{1}{1-\gamma}\left\|\frac{d_\rho(\pi^\star)}{\mu}\right\|_\infty \max_{\pi' \in \Pi}\left\langle \nabla V_\mu(\pi),\, \pi - \pi'\right\rangle$$

- **our improved convergence rate**

$$V_\rho(\pi^{(k)}) - V_\rho^\star \;\leq\; \frac{256|\mathcal{S}||\mathcal{A}|}{k\,(1-\gamma)^5}\left\|\frac{d_\rho(\pi^\star)}{\rho}\right\|_\infty^2$$

Weak gradient-mapping domination

consider minimizing $F(x) := f(x) + \Psi(x)$

- **proximal gradient method** with constant step size $\eta = \frac{1}{L}$

$$x^{k+1} = \arg\min_x \left\{ \left\langle \nabla f(x^k), x - x^k \right\rangle + \tfrac{L}{2}\|x - x^k\|^2 + \Psi(x) \right\}$$
$$= \mathbf{prox}_{\frac{1}{L}\Psi}\left(x^k - \tfrac{1}{L}\nabla f(x^k)\right)$$

- **Gradient mapping:** $G_L(x^k) = L\left(x^k - x^{k+1}\right)$
- **weak gradient mapping domination:** there exist $\omega > 0$ s.t.

$$\|G_L(x)\|_2 \geq \sqrt{2\omega}\big(F(x^+) - F^\star\big)$$

  KŁ exponent $= 1$, but with $F(x^+)$ instead of $F(x)$

- $O(1/k)$ **convergence to global optimum**

$$F(x^k) - F^\star \leq \max\left\{\tfrac{4L}{\omega k}, \left(\tfrac{\sqrt{2}}{2}\right)^k \left(F(x^0) - F^\star\right)\right\}$$

Policy mirror descent (PMD)

- **Bregman divergence**

$$D(p, p') := h(p) - h(p') - \langle \nabla h(p'), p - p' \rangle$$

weighted divergence: $D_\rho(\pi, \pi') := \mathbf{E}_{s\sim\rho}\big[D(\pi_s, \pi'_s)\big]$

- **general PMD**

$$\pi^{(k+1)} = \arg\min_{\pi\in\Pi} \left\{ \eta_k \big\langle \nabla V_\mu(\pi^{(k)}),\, \pi \big\rangle + \tfrac{1}{1-\gamma} D_{d_\mu(\pi^{(k)})}(\pi, \pi^{(k)}) \right\}$$

- **projected Q-descent**: for all $s \in \mathcal{S}$

$$\pi_s^{(k+1)} = \mathbf{proj}_{\Delta(\mathcal{A})}\left(\pi_s^{(k)} - \eta_k Q_s(\pi^{(k)})\right)$$

- **exponentiated Q-descent** (natural policy gradient)

$$\pi_{s,a}^{(k+1)} = \pi_{s,a}^{(k)} \frac{\exp\left(-\eta_k Q_{s,a}(\pi^{(k)})\right)}{z_s^{(k)}}$$

Convergence rates

- **sublinear rate** (constant step size) $\eta_k = \eta$

$$V_\rho(\pi^{(k)}) - V_\rho^\star \leq \frac{1}{k+1}\left(\frac{D_0^\star}{\eta(1-\gamma)} + \frac{1}{(1-\gamma)^2}\right)$$

- **linear rate** with **exponential step size:** $\eta_{k+1} \geq \eta_k/\gamma$

$$V_\rho(\pi^{(k)}) - V_\rho^\star \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^k \left(V_\rho(\pi^{(0)}) - V_\rho^\star + \frac{D_0^\star}{\eta_0\gamma}\right)$$

  where $\vartheta_\rho := \frac{1}{1-\gamma}\left\|\frac{d_\rho(\pi^\star)}{\rho}\right\|_\infty$ **(no entropy regularization)**

- **superlinear rate** if $\left\|d_\rho(\pi^\star)/d_\rho(\pi^{(k)})\right\|_\infty \to 1$
  - Convergence of transition probability matrix $P(\pi^k)$

$$\lim_{k\to\infty}\left\|P(\pi^k) - P(\pi^\star)\right\| = 0$$

  - exist $C > 0$ such that $\left\|P(\pi^k) - P(\pi^\star)\right\| \leq C\left(V_\rho(\pi^k) - V_\rho^\star\right)$

Connection with policy iteration

- **policy iteration (PI)**

$$\pi_s^{k+1} = \arg\min_{a \in \mathcal{A}} Q_{s,a}(\pi^k) = \arg\min_{p \in \Delta(\mathcal{A})} \left\langle Q_s(\pi^k),\ p \right\rangle$$

- **NPG (exponentiated Q-descent)**

$$\pi_s^{k+1} = \arg\min_{p \in \Delta(\mathcal{A})} \left\{ \eta_k \left\langle Q_s(\pi^k), p \right\rangle + D_{KL}(p, \pi_s^k) \right\}$$

equivalent to PI as $\eta_k \to \infty$

- **convergence rates**
  - policy iteration: $\|v(\pi^k) - V^\star\|_\infty \leq \frac{\gamma^k}{1-\gamma}$
  - NPG with $\eta_{k+1} = \eta_k/\gamma$

$$v_{\rho^\star}(\pi^k) - v_{\rho^\star}^\star \leq \gamma^k \left(\frac{1}{1-\gamma} + \frac{1}{\gamma \eta_0} \log |\mathcal{A}| \right)$$

Inexact PMD and sample complexity

$$\pi_s^{(k+1)} = \arg\min_{p \in \Delta(\mathcal{A})} \left\{ \eta_k \left\langle \widehat{Q}_s(\pi^{(k)}), p \right\rangle + D(p, \pi_s^{(k)}) \right\}$$

- **theorem:** if $\left\| \widehat{Q}(\pi^{(k)}) - Q(\pi^{(k)}) \right\|_\infty \leq \tau$ and $\eta_{k+1} > \eta_k/\gamma$, then

$$V_\rho(\pi^{(k)}) - V_\rho^\star \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^k \frac{2}{1-\gamma} + \frac{4\vartheta_\rho}{1-\gamma}\tau$$

- **sampling with generative model** (trajectory rollouts)
  - H: horizon of truncated trajectories
  - M: number of trajectories to average for each $\pi^{(k)}$ and (s, a)
  - if $M \geq \frac{\gamma^{-2H}}{2} \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)$, then with probability $\geq 1 - \delta$,

$$\left\| \widehat{Q}(\pi^{(k)}) - Q(\pi^{(k)}) \right\|_\infty \leq \frac{2\gamma^H}{1-\gamma}$$

  - overall sample complexity $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^8\epsilon^2} \left\| \frac{d_\rho(\pi^\star)}{\rho} \right\|_\infty^3\right)$