A general sample complexity analysis of vanilla policy gradient

Rui Yuan[†§], **Robert M. Gower**[§‡], **Alessandro Lazaric**[†]

[†]**Meta AI,** [§]**LTCI, Télécom Paris and** [‡]**Flatiron Institute**

https://arxiv.org/pdf/2107.11433.pdf

Accepted at AISTATS 2022

FACEBOOK AI

AI, machine learning, and optimization

- optimization: fundamental tool for AI and machine learning
- **importance of structure**
  - develop new algorithms and theory
    (natural gradient descent, mirror descent, variance reduction, . . .)
  - extend scope of fundamental algorithms and theory
    (also provide more insight of problem structure)
- *this talk*:
  - (stochastic) gradient decent
  - policy gradient methods

(Stochastic) gradient decent

Smooth nonconvex optimization

$\min_{x \in \mathbb{R}^d} f(x)$ with gradient descent $x^{k+1} = x^k - \eta_k \nabla f(x^k)$

Smooth nonconvex optimization

$\min_{x \in \mathbb{R}^d} f(x)$ with gradient descent $x^{k+1} = x^k - \eta_k \nabla f(x^k)$

- **smoothness:** the gradient is *Lipschitz*-continuous, i.e.

Smooth nonconvex optimization

$\min_{x \in \mathbb{R}^d} f(x)$ with gradient descent $x^{k+1} = x^k - \eta_k \nabla f(x^k)$

- **smoothness:** the gradient is *Lipschitz*-continuous, i.e.

$$\|\nabla f(x) - \nabla f(y)\|_2 \;\leq\; L\|x - y\|_2$$

Smooth nonconvex optimization

$\min_{x \in \mathbb{R}^d} f(x)$ with gradient descent $x^{k+1} = x^k - \eta_k \nabla f(x^k)$

- **smoothness:** the gradient is *Lipschitz*-continuous, i.e.

$$\|\nabla f(x) - \nabla f(y)\|_2 \;\leq\; L\|x - y\|_2$$

- **equivalent condition:** $\|\nabla^2 f(x)\| \;\leq\; L$
- **consequence:**

Smooth nonconvex optimization

$\min_{x \in \mathbb{R}^d} f(x)$ with gradient descent $x^{k+1} = x^k - \eta_k \nabla f(x^k)$

- **smoothness:** the gradient is *Lipschitz*-continuous, i.e.

$$\|\nabla f(x) - \nabla f(y)\|_2 \;\leq\; L\|x - y\|_2$$

- **equivalent condition:** $\|\nabla^2 f(x)\| \;\leq\; L$
- **consequence:**

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + (y - x)^\top \left(\int_0^1 (1 - t) \underbrace{\nabla^2 f(x + t(y - x))}_{\leq L} \, dt \right) (y - x)$$

Smooth nonconvex optimization

$\min_{x \in \mathbb{R}^d} f(x)$ with gradient descent $x^{k+1} = x^k - \eta_k \nabla f(x^k)$

- **smoothness:** the gradient is *Lipschitz*-continuous, i.e.

$$\|\nabla f(x) - \nabla f(y)\|_2 \;\leq\; L\|x - y\|_2$$

- **equivalent condition:** $\|\nabla^2 f(x)\| \;\leq\; L$
- **consequence:**

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + (y-x)^\top \left(\int_0^1 (1-t) \underbrace{\nabla^2 f(x + t(y-x))}_{\leq L} \, dt \right) (y - x)$$

$$\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2 := q(y)$$

Smooth nonconvex optimization

$\min_{x \in \mathbb{R}^d} f(x)$ with gradient descent $x^{k+1} = x^k - \eta_k \nabla f(x^k)$

- **smoothness:** the gradient is *Lipschitz*-continuous, i.e.

$$\|\nabla f(x) - \nabla f(y)\|_2 \;\leq\; L\|x - y\|_2$$

- **equivalent condition:** $\|\nabla^2 f(x)\| \;\leq\; L$
- **consequence:**

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + (y-x)^\top \left(\int_0^1 (1-t) \underbrace{\nabla^2 f(x + t(y-x))}_{\leq L} \, dt \right) (y - x)$$

$$\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2 := q(y)$$

- $y^* \in \arg\min q(y)$ s.t. $\nabla q(y^*) = 0$, that is,

$$\nabla f(x) + L(y^* - x) = 0 \iff y^* = x - \frac{1}{L}\nabla f(x),$$

$$f(y^*) \;\leq\; q(y^*) = f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2 \;\leq\; f(x)$$

Smooth nonconvex optimization

- **descent property:** with stepsize $\eta_k = 1/L$

Smooth nonconvex optimization

- **descent property:** with stepsize $\eta_k = 1/L$

$$
\begin{aligned}
f(x^K) - f(x^{K+1}) &\geq \frac{1}{2L}\|\nabla f(x^K)\|_2^2, \\
&\vdots \\
f(x^0) - f(x^1) &\geq \frac{1}{2L}\|\nabla f(x^0)\|_2^2. \\
&\Downarrow \\
f(x^0) - f(x^*) \geq f(x^0) - f(x^{K+1}) &\geq \frac{1}{2L}\sum_{k=0}^{K}\|\nabla f(x^k)\|_2^2.
\end{aligned}
$$

Smooth nonconvex optimization

- **descent property:** with stepsize $\eta_k = 1/L$

$$\begin{aligned} f(x^K) - f(x^{K+1}) &\geq \frac{1}{2L}\|\nabla f(x^K)\|_2^2, \\ &\vdots \\ f(x^0) - f(x^1) &\geq \frac{1}{2L}\|\nabla f(x^0)\|_2^2. \\ &\Downarrow \\ f(x^0) - f(x^*) \geq f(x^0) - f(x^{K+1}) &\geq \frac{1}{2L}\sum_{k=0}^{K}\|\nabla f(x^k)\|_2^2. \end{aligned}$$

- **sublinear convergence rate:** $\min_{0\leq k\leq K}\|\nabla f(x^k)\|_2^2 \leq \dfrac{2L(f(x^0) - f(x^*))}{K+1}$

  (1st-order stationary point)

Convergence to global optimum

- **gradient dominance** (Polyak-Lojasiewicz (PL) condition)

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \;\geq\; \mu\,(f(x) - f^*)$$

Convergence to global optimum

- **gradient dominance** (Polyak-Lojasiewicz (PL) condition)

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \;\geq\; \mu\,(f(x) - f^*)$$

  - gradient grows as quadratic function of sub-optimality

Convergence to global optimum

- **gradient dominance** (Polyak-Lojasiewicz (PL) condition)

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \;\geq\; \mu\,(f(x) - f^*)$$

  - gradient grows as quadratic function of sub-optimality
  - satisfied by strong convexity with convexity parameter μ

Convergence to global optimum

- **gradient dominance** (Polyak-Lojasiewicz (PL) condition)

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \;\geq\; \mu\,(f(x) - f^*)$$

  - gradient grows as quadratic function of sub-optimality
  - satisfied by strong convexity with convexity parameter μ
  - PL condition doesn't require uniqueness or convexity

Convergence to global optimum

- **gradient dominance** (Polyak-Lojasiewicz (PL) condition)

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \;\geq\; \mu\,(f(x) - f^*)$$

  - gradient grows as quadratic function of sub-optimality
  - satisfied by strong convexity with convexity parameter μ
  - PL condition doesn't require uniqueness or convexity
  - nonconvex f : guarantees convergence to global optimum (all stationary points are global optimum), e.g. $f(x) = x^2 + 3sin(x)^2$

Convergence to global optimum

- **gradient dominance** (Polyak-Lojasiewicz (PL) condition)

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \;\geq\; \mu\,(f(x) - f^*)$$

  - gradient grows as quadratic function of sub-optimality
  - satisfied by strong convexity with convexity parameter μ
  - PL condition doesn't require uniqueness or convexity
  - nonconvex f : guarantees convergence to global optimum (all stationary points are global optimum), e.g. $f(x) = x^2 + 3sin(x)^2$
- **linear convergence to global optimum:**

Convergence to global optimum

- **gradient dominance** (Polyak-Lojasiewicz (PL) condition)

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \;\geq\; \mu\,(f(x) - f^*)$$

  - gradient grows as quadratic function of sub-optimality
  - satisfied by strong convexity with convexity parameter μ
  - PL condition doesn't require uniqueness or convexity
  - nonconvex f : guarantees convergence to global optimum (all stationary points are global optimum), e.g. $f(x) = x^2 + 3sin(x)^2$
- **linear convergence to global optimum:**

$$f(x^K) - f(x^{K+1}) \;\geq\; \frac{1}{2L}\|\nabla f(x^K)\|_2^2 \;\geq\; \frac{\mu}{L}(f(x^K) - f^*)$$

Convergence to global optimum

- **gradient dominance** (Polyak-Lojasiewicz (PL) condition)

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \;\geq\; \mu\,(f(x) - f^*)$$

  - gradient grows as quadratic function of sub-optimality
  - satisfied by strong convexity with convexity parameter μ
  - PL condition doesn't require uniqueness or convexity
  - nonconvex f : guarantees convergence to global optimum (all stationary points are global optimum), e.g. $f(x) = x^2 + 3sin(x)^2$

- **linear convergence to global optimum:**

$$f(x^K) - f(x^{K+1}) \;\geq\; \frac{1}{2L}\|\nabla f(x^K)\|_2^2 \;\geq\; \frac{\mu}{L}(f(x^K) - f^*)$$

$$\implies f(x^{K+1}) - f^* \;\leq\; \left(1 - \frac{\mu}{L}\right)(f(x^K) - f^*)$$

Convergence to global optimum

- **gradient dominance** (Polyak-Lojasiewicz (PL) condition)

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \;\geq\; \mu\,(f(x) - f^*)$$

  - gradient grows as quadratic function of sub-optimality
  - satisfied by strong convexity with convexity parameter μ
  - PL condition doesn't require uniqueness or convexity
  - nonconvex f : guarantees convergence to global optimum (all stationary points are global optimum), e.g. $f(x) = x^2 + 3sin(x)^2$

- **linear convergence to global optimum:**

$$f(x^K) - f(x^{K+1}) \;\geq\; \frac{1}{2L}\|\nabla f(x^K)\|_2^2 \;\geq\; \frac{\mu}{L}(f(x^K) - f^*)$$

$$\implies f(x^{K+1}) - f^* \;\leq\; \left(1 - \frac{\mu}{L}\right)(f(x^K) - f^*)$$

$$\implies f(x^{K+1}) - f^* \;\leq\; \left(1 - \frac{\mu}{L}\right)^K (f(x^0) - f^*)$$

Convergence to global optimum

- **weak gradient dominance** (weak PL condition)

$$\|\nabla f(x)\|_2 \;\geq\; \sqrt{2\mu}\,(f(x) - f^*)$$

Convergence to global optimum

- **weak gradient dominance** (weak PL condition)

$$\|\nabla f(x)\|_2 \;\geq\; \sqrt{2\mu}\,(f(x) - f^*)$$

- combined with smooth descent property:

$$f(x^K) - f(x^{K+1}) \;\geq\; \frac{1}{2L}\|\nabla f(x^K)\|_2^2 \;\geq\; \frac{\mu}{L}(f(x^K) - f^*)^2$$

Convergence to global optimum

- **weak gradient dominance** (weak PL condition)

$$\|\nabla f(x)\|_2 \;\geq\; \sqrt{2\mu}\,(f(x) - f^*)$$

- combined with smooth descent property:

$$f(x^K) - f(x^{K+1}) \;\geq\; \frac{1}{2L}\|\nabla f(x^K)\|_2^2 \;\geq\; \frac{\mu}{L}(f(x^K) - f^*)^2$$

- $\mathcal{O}(1/K)$ **sublinear convergence to global optimum:**

$$f(x^K) - f^* \;\leq\; \frac{f(x^0) - f^*}{1 + K \cdot \frac{\mu}{L}(f(x^0) - f^*)}$$

Proof of $\mathcal{O}(1/K)$ convergence

- let $\delta_k = f(x^k) - f^*$, then

$$\delta_k - \delta_{k+1} \;\geq\; \frac{\mu}{L}\delta_k^2$$

Proof of $\mathcal{O}(1/K)$ convergence

- let $\delta_k = f(x^k) - f^*$, then

$$\delta_k - \delta_{k+1} \;\geq\; \frac{\mu}{L}\delta_k^2$$

- dividing both sides by $\delta_k \delta_{k+1}$ and using $\delta_k \geq \delta_{k+1}$:

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} \geq \frac{\mu}{L}\frac{\delta_k}{\delta_{k+1}} \geq \frac{\mu}{L}$$

Proof of $\mathcal{O}(1/K)$ convergence

- let $\delta_k = f(x^k) - f^*$, then

$$\delta_k - \delta_{k+1} \;\geq\; \frac{\mu}{L}\delta_k^2$$

- dividing both sides by $\delta_k \delta_{k+1}$ and using $\delta_k \geq \delta_{k+1}$:

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} \geq \frac{\mu}{L}\frac{\delta_k}{\delta_{k+1}} \geq \frac{\mu}{L}$$

- telescoping sum over iterations $0, 1, \cdots, k-1$:

$$\frac{1}{\delta_{k+1}} - \frac{1}{\delta_0} \geq k \cdot \frac{\mu}{L} \quad \Longrightarrow \quad \delta_k \leq \frac{1}{\frac{1}{\delta_0} + k \cdot \frac{\mu}{L}}$$

Summary of (stochastic) gradient descent

$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}_\zeta \left[f_\zeta(x) \right]$ with stochastic gradient descent (SGD)

$x^{k+1} = x^k - \eta_k \nabla f_\zeta(x^k)$

assumptions	gradient descent	SGD	criteria
`smooth + ABC`	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-4})$	$\min_{0 \le k \le K} \mathbb{E}_\zeta \left[\|\nabla f_\zeta(x^k)\|_2^2 \right] \le \epsilon^2$
`smooth + ABC + PL`	linear	$\mathcal{O}(\epsilon^{-1})$	$f - f^* \le \epsilon$
`smooth + ABC + weak PL`	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-3})$	$f - f^* \le \epsilon$

- smooth: $\|\nabla^2 f(x)\| \le L$
- ABC: $\mathbb{E}_\zeta \left[\|\nabla f_\zeta(x)\|_2^2 \right] \le 2A(f(x) - f^*) + B\|\nabla f(x)\|_2^2 + C$
- PL: $\frac{1}{2}\|\nabla f(x)\|_2^2 \ge \mu \left(f(x) - f^* \right)$
- weak PL: $\|\nabla f(x)\|_2 \ge \sqrt{2\mu} \left(f(x) - f^* \right)$

Policy gradient methods

Context

- Consider a classic Markov decision process (MDP)

$$\mathcal{M} \;=\; \{\mathcal{S},\ \mathcal{A},\ \mathcal{P},\ \mathcal{R},\ \gamma,\ \rho\}$$

Markov property: $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$

Context

- Consider a classic Markov decision process (MDP)

$$\mathcal{M} \quad = \quad \{\mathcal{S},\ \mathcal{A},\ \mathcal{P},\ \mathcal{R},\ \gamma,\ \rho\}$$

  Markov property: $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$

- Total discounted reward provided by trajectory $\tau = (s_0, a_0, s_1, a_1, \cdots)$

$$\mathcal{R}(\tau) \quad = \quad \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)$$

Context

- Consider a classic Markov decision process (MDP)

$$\mathcal{M} \quad = \quad \{\mathcal{S},\ \mathcal{A},\ \mathcal{P},\ \mathcal{R},\ \gamma,\ \rho\}$$

  Markov property: $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$
- Total discounted reward provided by trajectory $\tau = (s_0, a_0, s_1, a_1, \cdots)$

$$\mathcal{R}(\tau) \quad = \quad \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)$$

- *Objective*: Solving an MDP $\mathcal{M}$, i.e. finding the optimum parametric policy $\pi_\theta : s \in \mathcal{S} \to \Delta(\mathcal{A})$

$$\max_\theta J(\theta) \ \stackrel{\text{def}}{=}\ \mathbb{E}_{\tau \sim p(\cdot|\pi_\theta, \mathcal{M})} \left[\mathcal{R}(\tau)\right] \ \stackrel{\text{def}}{=}\ \mathbb{E}_{\tau \sim p(\cdot|\theta)} \left[\mathcal{R}(\tau)\right]$$

Context

- Consider a classic Markov decision process (MDP)

$$\mathcal{M} \quad = \quad \{\mathcal{S},\ \mathcal{A},\ \mathcal{P},\ \mathcal{R},\ \gamma,\ \rho\}$$

  Markov property: $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$
- Total discounted reward provided by trajectory $\tau = (s_0, a_0, s_1, a_1, \cdots)$

$$\mathcal{R}(\tau) \quad = \quad \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)$$

- *Objective*: Solving an MDP $\mathcal{M}$, i.e. finding the optimum parametric policy $\pi_\theta : s \in \mathcal{S} \to \Delta(\mathcal{A})$

$$\max_\theta J(\theta) \ \stackrel{\text{def}}{=}\ \mathbb{E}_{\tau \sim p(\cdot|\pi_\theta, \mathcal{M})} \left[\mathcal{R}(\tau)\right] \ \stackrel{\text{def}}{=}\ \mathbb{E}_{\tau \sim p(\cdot|\theta)} \left[\mathcal{R}(\tau)\right]$$

- From the Markov property, we know that for $\tau = (s_0, a_0, s_1, a_1, \cdots)$, we have

$$p(\tau \mid \theta) = \rho(s_0) \prod_{t=0}^{\infty} \pi_\theta(a_t \mid s_t) \mathcal{P}(s_{t+1} \mid s_t, a_t)$$

Policy gradient methods as SGD

Policy gradient (PG) methods (REINFORCE [Williams, 1992], GPOMDP [Sutton et al., 2000, Baxter and Bartlett, 2001])

$$\theta_{k+1} \quad = \quad \theta_k + \eta_k \widehat{\nabla_\theta J(\theta_k)}$$

Policy gradient methods as SGD

Policy gradient (PG) methods (REINFORCE [Williams, 1992], GPOMDP [Sutton et al., 2000, Baxter and Bartlett, 2001])

$$\theta_{k+1} \quad = \quad \theta_k + \eta_k \widehat{\nabla_\theta J(\theta_k)}$$

Compute $\nabla_\theta J(\theta)$

Policy gradient methods as SGD

Policy gradient (PG) methods (REINFORCE [Williams, 1992], GPOMDP [Sutton et al., 2000, Baxter and Bartlett, 2001])

$$\theta_{k+1} \quad = \quad \theta_k + \eta_k \widehat{\nabla_\theta J(\theta_k)}$$

Compute $\nabla_\theta J(\theta)$

$$\nabla_\theta J(\theta) \quad = \quad \nabla_\theta \mathbb{E}_{\tau \sim p(\cdot|\theta)} \left[\mathcal{R}(\tau) \right]$$

Policy gradient methods as SGD

Policy gradient (PG) methods (REINFORCE [Williams, 1992], GPOMDP [Sutton et al., 2000, Baxter and Bartlett, 2001])

$$\theta_{k+1} = \theta_k + \eta_k \widehat{\nabla_\theta J(\theta_k)}$$

Compute $\nabla_\theta J(\theta)$

$$\begin{aligned} \nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_{\tau \sim p(\cdot|\theta)} \left[\mathcal{R}(\tau) \right] \\ &= \int \nabla_\theta p(\tau \mid \theta) \mathcal{R}(\tau) d\tau \end{aligned}$$

Policy gradient methods as SGD

Policy gradient (PG) methods (REINFORCE [Williams, 1992], GPOMDP [Sutton et al., 2000, Baxter and Bartlett, 2001])

$$\theta_{k+1} = \theta_k + \eta_k \widehat{\nabla_\theta J(\theta_k)}$$

Compute $\nabla_\theta J(\theta)$

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_{\tau \sim p(\cdot|\theta)} \left[\mathcal{R}(\tau) \right] \\
&= \int \nabla_\theta p(\tau \mid \theta) \mathcal{R}(\tau) d\tau \\
&= \int p(\tau \mid \theta) \nabla_\theta \log p(\tau \mid \theta) \mathcal{R}(\tau) d\tau
\end{aligned}$$

Policy gradient methods as SGD

Policy gradient (PG) methods (REINFORCE [Williams, 1992], GPOMDP [Sutton et al., 2000, Baxter and Bartlett, 2001])

$$\theta_{k+1} = \theta_k + \eta_k \widehat{\nabla_\theta J(\theta_k)}$$

Compute $\nabla_\theta J(\theta)$

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_{\tau \sim p(\cdot|\theta)} \left[\mathcal{R}(\tau) \right] \\
&= \int \nabla_\theta p(\tau \mid \theta) \mathcal{R}(\tau) d\tau \\
&= \int p(\tau \mid \theta) \nabla_\theta \log p(\tau \mid \theta) \mathcal{R}(\tau) d\tau \\
&= \mathbb{E}_{\tau \sim p(\cdot|\theta)} \left[\nabla_\theta \log p(\tau \mid \theta) \mathcal{R}(\tau) \right]
\end{aligned}$$

Policy gradient methods as SGD

Policy gradient (PG) methods (REINFORCE [Williams, 1992], GPOMDP [Sutton et al., 2000, Baxter and Bartlett, 2001])

$$\theta_{k+1} = \theta_k + \eta_k \widehat{\nabla_\theta J(\theta_k)}$$

Compute $\nabla_\theta J(\theta)$

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_{\tau \sim p(\cdot|\theta)} [\mathcal{R}(\tau)] \\
&= \int \nabla_\theta p(\tau \mid \theta) \mathcal{R}(\tau) d\tau \\
&= \int p(\tau \mid \theta) \nabla_\theta \log p(\tau \mid \theta) \mathcal{R}(\tau) d\tau \\
&= \mathbb{E}_{\tau \sim p(\cdot|\theta)} [\nabla_\theta \log p(\tau \mid \theta) \mathcal{R}(\tau)] \\
\nabla_\theta \log p(\tau \mid \theta) &= \nabla_\theta \log \left(\rho(s_0) \prod_{t=0}^{\infty} \pi_\theta(a_t \mid s_t) \mathcal{P}(s_{t+1} \mid s_t, a_t) \right) \\
&= \sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t \mid s_t)
\end{aligned}$$

Policy gradient methods as SGD

Compute $\nabla J(\theta)$

$$\nabla J(\theta) \quad = \quad \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \sum_{t'=0}^{\infty} \nabla_\theta \log \pi_\theta(a_{t'} \mid s_{t'}) \right]$$

Policy gradient methods as SGD

Compute $\nabla J(\theta)$

$$\nabla J(\theta) \quad = \quad \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \sum_{t'=0}^{\infty} \nabla_\theta \log \pi_\theta(a_{t'} \mid s_{t'}) \right]$$

Compute an empirical estimator of the gradient by sampling m *truncated* trajectories $\tau = (s_0, a_0, s_1, a_1, \cdots, s_H, a_H)$

$$\widehat{\nabla}_m J(\theta) \quad = \quad \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t^i, a_t^i) \cdot \sum_{t'=0}^{H-1} \nabla_\theta \log \pi_\theta(a_{t'}^i \mid s_{t'}^i)$$

Policy gradient methods as SGD

Compute $\nabla J(\theta)$

$$\nabla J(\theta) \quad = \quad \mathbb{E}_\tau \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \sum_{t'=0}^{\infty} \nabla_\theta \log \pi_\theta(a_{t'} \mid s_{t'}) \right]$$

Compute an empirical estimator of the gradient by sampling m *truncated* trajectories $\tau = (s_0, a_0, s_1, a_1, \cdots, s_H, a_H)$

$$\widehat{\nabla}_m J(\theta) \quad = \quad \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t^i, a_t^i) \cdot \sum_{t'=0}^{H-1} \nabla_\theta \log \pi_\theta(a_{t'}^i \mid s_{t'}^i)$$

Vanilla policy gradient (REINFORCE [Williams, 1992], GPOMDP [Sutton et al., 2000, Baxter and Bartlett, 2001])

$$\theta_{k+1} \quad = \quad \theta_k + \eta \widehat{\nabla}_m J(\theta_k)$$

Current literatures of vanilla PG: *fragmentary* !

- Exact PG [Fazel et al., 2018, Agarwal et al., 2021, Zhang et al., 2020a, Mei et al., 2020] vs stochastic PG [Papini et al., 2019, Liu et al., 2020, Zhang et al., 2020c, Xiong et al., 2021]
- Different criteria of the convergence results: first-order stationary point [Papini et al., 2019, Zhang et al., 2020c]; global optimum [Fazel et al., 2018, Agarwal et al., 2021, Zhang et al., 2020a, Mei et al., 2020]; average regret to the global optimum [Zhang et al., 2020b, Liu et al., 2020]
- Different RL settings: linear quadratic regulator [Fazel et al., 2018], softmax tabular policy w/o different regularizations [Agarwal et al., 2021, Zhang et al., 2020a,b, Mei et al., 2020], Fisher-non-degenerate policy [Liu et al., 2020, Ding et al., 2021]
- Different assumptions: Lipschitz and smooth policy [Liu et al., 2020, Zhang et al., 2020c, Xiong et al., 2021], bijection between the primal and the dual space [Zhang et al., 2020a]
- Large mini-batch (e.g. $\mathcal{O}(\epsilon^{-1})$, $\mathcal{O}(\epsilon^{-2})$) per iteration for stochastic updates [Papini et al., 2019, Liu et al., 2020, Zhang et al., 2020c, Xiong et al., 2021]

Contribution

- We propose a general PG analysis with weaker assumptions. The generality of our assumption allows us to unify much of the fragmented results in the literature under one guise without lost of the performance.
- Recover existing $\widetilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity guarantees with weaker assumptions for *wider ranges* of parameters (e.g. mini-batch m from 1 to $\mathcal{O}(\epsilon^{-2})$)
- New $\widetilde{\mathcal{O}}(\epsilon^{-3})$ sample complexity for *global optimum* guarantees with additional relaxed weak gradient domination assumption, including *Fisher-non-degenerate parametrized policies* as special case

Assumptions

Assumption (Smoothness)

There exists $L > 0$ such that, for all $\theta, \theta' \in \mathbb{R}^d$, we have

$$\left|J(\theta') - J(\theta) - \left\langle \nabla J(\theta), \theta' - \theta \right\rangle\right| \leq \frac{L}{2}\|\theta' - \theta\|^2. \qquad \text{(smoothness)}$$

Assumption (ABC [Khaled and Richtárik, 2020])

The stochastic gradient satisfies

$$\mathbb{E}\left[\|\widehat{\nabla}_m J(\theta)\|^2\right] \leq 2A(J^* - J(\theta)) + B\|\nabla J_H(\theta)\|^2 + C, \qquad \text{(ABC)}$$

for some $A, B, C \geq 0$ and all $\theta \in \mathbb{R}^d$.

Here $J_H(\theta) = \mathbb{E}_\tau\left[\sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t, a_t)\right]$ is the expected truncated total reward function.

Convergence result under ABC assumption

Theorem

Suppose that Assumption smoothness and ABC are satisfied. We choose a constant stepsize η such that $\eta \in \left(0, \frac{2}{LB}\right)$ where B can be zero.[a] *Let $\delta_0 \overset{def}{=} J^* - J(\theta_0)$. If $A > 0$, then PG satisfies*

$$\min_{0 \le t \le T-1} \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \le \frac{2\delta_0(1 + L\eta^2 A)^T}{\eta T(2 - LB\eta)} + \frac{LC\eta}{2 - LB\eta} + \mathcal{O}(\gamma^H).$$

If $A = 0$, we have

$$\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] \le \frac{2\delta_0}{\eta T(2 - LB\eta)} + \frac{LC\eta}{2 - LB\eta} + \mathcal{O}(\gamma^H),$$

where θ_U is uniformly sampled from $\{\theta_0, \theta_1, \cdots, \theta_{T-1}\}$.

[a] We set $\frac{1}{0} = \infty$.

Sample complexity under ABC assumption

If we set the parameters as

$$\eta = \min\left\{\frac{1}{\sqrt{LAT}}, \frac{1}{LB}, \frac{\epsilon}{2LC}\right\},$$
$$T \geq \frac{12\delta_0 L}{\epsilon^2}\max\left\{B, \frac{12\delta_0 A}{\epsilon^2}, \frac{2C}{\epsilon^2}\right\},$$
$$H = \mathcal{O}(\log \epsilon^{-1}),$$

then $\min_{0\leq t\leq T-1} \mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] = \mathcal{O}(\epsilon^{-2})$.

☞ Sample complexity (i.e., *single step interaction with the environment and single sampled trajectory* (s, a) *per iteration*): $TH = \widetilde{\mathcal{O}}(\epsilon^{-4})$

☞ For the full exact gradient ($A = C = 0, B = 1$): $T = \mathcal{O}(\epsilon^{-2})$

Global optimum convergence under relaxed weak gradient domination assumption

Assumption (Relaxed weak gradient domination)

We say that J satisfies the relaxed weak gradient domination condition if for all $\theta \in \mathcal{R}^d$, there exists $\mu > 0$ and $\epsilon' \geq 0$ such that

$$\epsilon' + \|\nabla J_H(\theta)\| \;\geq\; 2\sqrt{\mu}\,(J^* - J(\theta))\,. \qquad \text{(weak PL)}$$

Theorem

Suppose that Assumption smoothness, ABC and weak PL hold. Given $\epsilon > 0$, choose the stepsize $\eta_t = \mathcal{O}(t^{-1})$ when $t > t_0$ with $t_0 = [T/2]$, let the horizon $H = \mathcal{O}(\log \epsilon^{-1})$. If $\epsilon' = 0$, we choose the number of iterations $T = \mathcal{O}(\epsilon^{-3})$; if $\epsilon' > 0$, we choose $T = \mathcal{O}((\epsilon')^{-2}\epsilon^{-1})$. Then

$$\min_{t \in \{0,1,\cdots,T\}} J^* - \mathbb{E}J(\theta_t) \leq \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon').$$

Applications

Reinforcement Learning Assumption

Assumption ((Expected) Lipschitz and smooth policy (E-LS) [Papini et al., 2019])

There exists constants $G, F > 0$ such that for every action $a \in \mathcal{A}$ and every state $s \in \mathcal{S}$, the gradient and Hessian of $\log \pi_\theta(a \mid s)$ satisfy

$$\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[\|\nabla_\theta \log \pi_\theta(a \mid s)\|^2 \right] \leq G^2,$$
$$\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[\|\nabla^2_\theta \log \pi_\theta(a \mid s)\| \right] \leq F.$$

Reinforcement Learning Assumption

Assumption ((Expected) Lipschitz and smooth policy (E-LS) [Papini et al., 2019])

There exists constants $G, F > 0$ such that for every action $a \in \mathcal{A}$ and every state $s \in \mathcal{S}$, the gradient and Hessian of $\log \pi_\theta(a \mid s)$ satisfy

$$\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[\|\nabla_\theta \log \pi_\theta(a \mid s)\|^2 \right] \leq G^2,$$

$$\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[\|\nabla^2_\theta \log \pi_\theta(a \mid s)\| \right] \leq F.$$

Softmax policy (w/o regularization (e.g., entropy, log barrier)) and *Gaussian policy* with *unbounded* action space satisfy this assumption.

Reinforcement Learning Assumption

Assumption ((Expected) Lipschitz and smooth policy (E-LS) [Papini et al., 2019])

There exists constants $G, F > 0$ such that for every action $a \in \mathcal{A}$ and every state $s \in \mathcal{S}$, the gradient and Hessian of $\log \pi_\theta(a \mid s)$ satisfy

$$\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[\|\nabla_\theta \log \pi_\theta(a \mid s)\|^2 \right] \leq G^2,$$
$$\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[\|\nabla^2_\theta \log \pi_\theta(a \mid s)\| \right] \leq F.$$

Softmax policy (w/o regularization (e.g., entropy, log barrier)) and *Gaussian policy* with *unbounded* action space satisfy this assumption.

$$\text{Softmax policy} \quad \pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}$$

$$\text{Guassian policy} \quad \pi_\theta(a \mid s) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{a - \theta^\top \phi(s)}{\sigma} \right)^2 \right\}$$

Sufficient conditions for Asm. smoothness and ABC (1/2)

Lemma

Under Asm. E-LS, $J(\cdot)$ is L-smooth, namely $\|\nabla^2 J(\theta)\| \leq L$ for all θ which is a sufficient condition of Asm. smoothness, with

$$L = \frac{\mathcal{R}_{\max}}{(1-\gamma)^2}(G^2 + F). \tag{1}$$

Here we assume $\mathcal{R} \in [-\mathcal{R}_{\max}, \mathcal{R}_{\max}]$.

Sufficient conditions for Asm. smoothness and ABC (2/2)

Lemma (Bounded variance of the gradient estimator)

Under Asm. E-LS, Asm. ABC holds with $A = 0$*, that is,*

$$\mathbb{E}\left[\|\widehat{\nabla}_m J(\theta)\|^2\right] \leq \underbrace{\left(1 - \frac{1}{m}\right)}_{=B} \|\nabla J_H(\theta)\|^2 + \underbrace{\frac{\nu}{m}}_{=C},$$

where m *is the mini-batch size, and* $\nu = \dfrac{HG^2\mathcal{R}_{\max}^2}{(1-\gamma)^2}$ *for REINFORCE or* $\nu = \dfrac{G^2\mathcal{R}_{\max}^2}{(1-\gamma)^3}$ *for GPOMDP.*

Convergence of policy gradient

Corollary

Suppose that Assumption E-LS is satisfied. Let $\delta_0 \stackrel{def}{=} J^* - J(\theta_0)$. *Any PG method with a mini-batch sampling of size* m *and stepsize*

$$\eta \in \left(0, \frac{2}{L(1-1/m)}\right),$$

we have

$$\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] \leq \frac{2\delta_0}{\eta T\left(2 - L\eta\left(1-\frac{1}{m}\right)\right)} + \frac{L\nu\eta}{m\left(2 - L\eta\left(1-\frac{1}{m}\right)\right)} + \mathcal{O}(\gamma^H).$$

Sample complexity of policy gradient

If we set the parameters as

$$
\begin{aligned}
m &\in \left[1,\ \frac{2\nu}{\epsilon^2}\right], \\
T \text{ s.t. } Tm &\geq \frac{8\delta_0 L\nu}{\epsilon^4}, \\
\eta &= \frac{\epsilon^2 m}{2L\nu}, \\
H &= \mathcal{O}((1-\gamma)^{-1}\log\epsilon^{-1}),
\end{aligned}
$$

then $\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] = \mathcal{O}(\epsilon^{-2})$.

👉 Sample complexity: $TmH = \widetilde{\mathcal{O}}((1-\gamma)^{-6}\epsilon^{-4})$

Fisher-non-degenerate parametrized policy (1/2)

Assumption (Fisher-non-degenerate, Asm. 2.1 in Ding et al. [2021])

For all $\theta \in \mathcal{R}^d$, there exists $\mu_F > 0$ s.t. the Fisher information matrix $F_\rho(\theta)$ induced by policy π_θ and initial state distribution ρ satisfies

$$F_\rho(\theta) \overset{def}{=} \mathbb{E}_{(s,a)\sim v_\rho^{\pi_\theta}} \left[\nabla_\theta \log \pi_\theta(a \mid s) \nabla_\theta \log \pi_\theta(a \mid s)^\top \right] \geq \mu_F \mathbf{I}_d, \tag{FI}$$

where $v_\rho^{\pi_\theta}$ is the state-action visitation measure defined as

$$v_\rho^{\pi_\theta}(s,a) \overset{def}{=} (1-\gamma)\mathbb{E}_{s_0\sim\rho} \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a | s_0, \pi_\theta).$$

Fisher-non-degenerate parametrized policy (2/2)

Assumption (Compatible, Asm. 4.6 in Ding et al. [2021])

For all $\theta \in \mathcal{R}^d$, there exists $\epsilon_{bias} > 0$ s.t. the transferred compatible function approximation error with $(s,a) \sim v_\rho^{\pi_{\theta^*}}$ satisfies

$$\mathbb{E}(A^{\pi_\theta}(s,a) - (1-\gamma)u^{*\top}\nabla_\theta \pi_\theta(a \mid s))^2 \leq \epsilon_{bias}, \qquad \text{(compatible)}$$

where $v_\rho^{\pi_{\theta^*}}$ is the state-action distribution induced by an optimal policy π_{θ^*}, $u^* = (F_\rho(\theta))^\dagger \nabla J(\theta)$ is the natural policy gradient update.

Global convergence of Fisher-non-degenerate parametrized policy

Proposition (Lemma 4.7 in Ding et al. [2021])

If the policy π_θ satisfies Assumption E-LS, FI and compatible, then

$$\frac{\mu_F\sqrt{\epsilon_{bias}}}{(1-\gamma)G} + \|\nabla J_H(\theta)\| \geq \frac{\mu_F}{G}(J^* - J(\theta)).$$

Corollary

If the policy π_θ satisfies Asm. E-LS, FI and compatible, consider the setting of Thm. 5 with $\epsilon' = \frac{\mu_F\sqrt{\epsilon_{bias}}}{(1-\gamma)G}$ and $\mu = \frac{\mu_F^2}{4G^2}$. Then $\min_{t\in\{0,1,\cdots,T\}} J^* - \mathbb{E}J(\theta_t) \leq \mathcal{O}(\epsilon) + \mathcal{O}(\sqrt{\epsilon_{bias}})$ and the sample complexity $T \times H = \widetilde{\mathcal{O}}(\epsilon^{-3})$ when $\epsilon_{bias} = 0$ or $T \times H = \widetilde{\mathcal{O}}((\epsilon_{bias}\cdot\epsilon)^{-1})$ when $\epsilon_{bias} > 0$.

Summary of (stochastic) policy gradient

$$\max_{\theta\in\mathbb{R}^n} J(\theta) := \mathbb{E}_{\tau\sim p(\cdot|\pi_\theta,\mathcal{M})}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)\right] \text{ with (stochastic) policy gradient (PG)}$$

$$\theta_{k+1} = \theta_k + \eta\widehat{\nabla}_m J(\theta_k)$$

assumptions	exact PG	stochastic PG	criteria
`smooth + ABC`	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-4})$	$\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] = \mathcal{O}(\epsilon^{-2})$
`smooth + ABC + PL`	linear [Mei et al., 2020]	$\mathcal{O}(\epsilon^{-1})$	$J^* - J \leq \epsilon$
`smooth + ABC + weak PL`	$\mathcal{O}(\epsilon^{-1})$ [Mei et al., 2020]	$\mathcal{O}(\epsilon^{-3})$	$J^* - J \leq \epsilon$

- smooth: $\|\nabla^2 J(\theta)\| \leq L$
- ABC: $\mathbb{E}\left[\|\widehat{\nabla}_m J(\theta)\|^2\right] \leq 2A(J^* - J(\theta)) + B\|\nabla J_H(\theta)\|^2 + C$
- PL: $\frac{1}{2}\|\nabla J_H(\theta)\|_2^2 \geq \mu\,(J^* - J(\theta))$
- weak PL: $\|\nabla J_H(\theta)\|_2 + \epsilon' \geq \sqrt{2\mu}\,(J^* - J(\theta))$

Other variants of policy gradient methods

- stochastic gradient descent -> *policy gradient* -> *actor-critic* [Konda and Tsitsiklis, 2000]
- natural gradient descent -> *natural policy gradient* [Kakade, 2002] -> *TRPO* [Schulman et al., 2015] and *PPO* [Schulman et al., 2017]
- stochastic mirror descent -> *policy mirror descent* -> *mirror descent policy optimization (MDPO)* [Tomar et al., 2021]
- variance reduced gradient methods -> *variance reduced policy gradient methods* [Papini et al., 2018, Xu et al., 2020, Huang et al., 2020]

Future work

- Extend the ABC assumption and its analysis to other RL settings (e.g., linear quadratic regulator)
- Extend the ABC assumption and its analysis to other methods (e.g., projected PG, policy mirror descent, natural policy gradient)
- Consider improved convergence analysis for variance reduced methods with the relaxed weak gradient domination assumption

Thank you

FACEBOOK AI

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98): 1–76, 2021.

J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, Nov 2001. ISSN 1076-9757. doi: $10.1613/\text{jair}.806$.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global convergence of momentum-based policy gradient, 2021.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1467–1476. PMLR, 10–15 Jul 2018.

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods, 2020.

Sham M Kakade. A natural policy gradient. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.

Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world, 2020.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7624–7636. Curran Associates, Inc., 2020.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6820–6829. PMLR, 13–18 Jul 2020.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4026–4035. PMLR, 2018.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1889–1897, Lille, France, 07–09 Jul 2015. PMLR.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Richard S Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1057–1063. MIT Press, 2000.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization, 2021.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Huaqing Xiong, Tengyu Xu, Yingbin Liang, and Wei Zhang. Non-asymptotic convergence of adam-type reinforcement learning algorithms under markovian sampling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10460–10468, May 2021.

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations*, 2020.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4572–4583. Curran Associates, Inc., 2020a.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with reinforce, 2020b.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020c. doi: 10.1137/19M1288012.