Convergence and Iteration Complexity of Policy Gradient Method for Infinite-Horizon Reinforcement Learning

Kaiqing Zhang⋆ Alec Koppel§ Hao Zhu‡ Tamer Başar⋆
⋆UIUC ‡UT Austin §U.S. Army Research Laboratory

Large-Scale Distributed Optimization and Decentralized Control I
IEEE Conference on Decision and Control

Dec. 13, 2019

Reinforcement Learning

Reinforcement learning: data-driven control

$\Rightarrow$ unknown system model/cost function

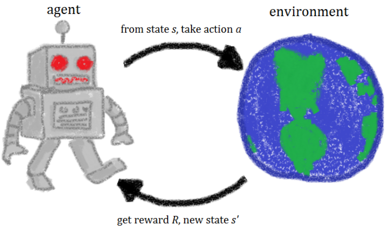
$\Rightarrow$ parameterize policy/cost as stat. model for high dimensional spaces

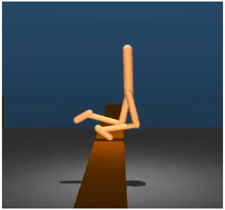
Recent successes:

$\Rightarrow$ AlphaGo Zero [Silver et al. '17]

$\Rightarrow$ Bipedal walker on terrain [Heess et al. '17]

$\Rightarrow$ Personalized web services [Theocharous et al. '15]





Problem Formulation

Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma)$

$\Rightarrow$ State space $\mathcal{S}$, action space $\mathcal{A}$ (high-dim. or even continuous)

$\Rightarrow$ Markov transition kernel $\mathbb{P}(s' \mid s, a) : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$

$\Rightarrow$ Reward $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor $\gamma \in (0, 1)$

Stochastic policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, i.e., $a_t \sim \pi(\cdot \mid s_t)$

Infinite-horizon setting value function:

$$V(s) = \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t \cdot R(s_t, a_t) \,\middle|\, s_0 = s \right),$$

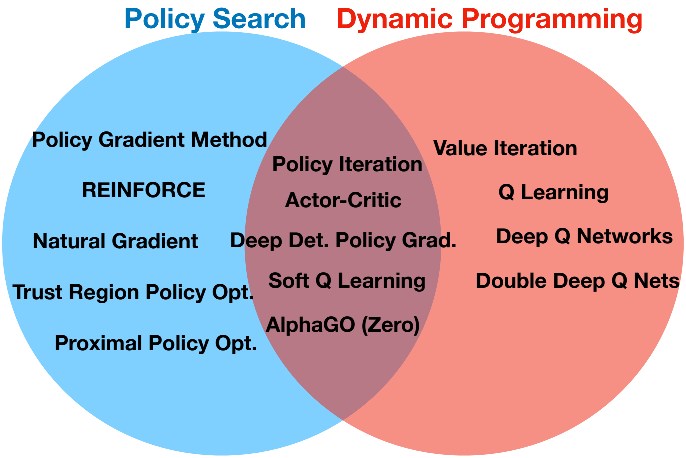
Goal: find $\{a_t = \pi(s_t)\}$ to maximize $V_\pi(s) := \mathbb{E}[V(s) \mid a \sim \pi(s)]$

$\max_{\pi \in \Pi} V_\pi(s)$ where Π is some family of distributions

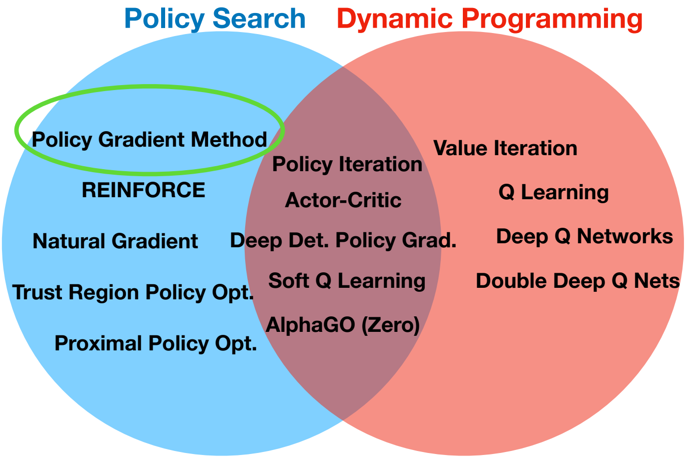
$\Rightarrow$ E.g., Gaussian $\pi = \pi_\theta$ w/ $\theta \in \mathbb{R}^d$ $\Rightarrow$ $\pi_\theta(\cdot \mid s) = \mathcal{N}(\phi(s)^\top \theta, \sigma^2)$

$\Rightarrow$ Define action-state value (Q) function $Q_\pi(s, a) = \mathbb{E}[V_\pi(s) \mid a_0 = a]$

Literature Landscape



Literature Landscape



Context

Pros of policy gradient [Silver ′14]:

- Better convergence properties
- Effective in high-dimensional or continuous action spaces
- Can learn stochastic policies

Cons of policy gradient [Silver ′14]:

- Typically converge to a local rather than global optimum

Context

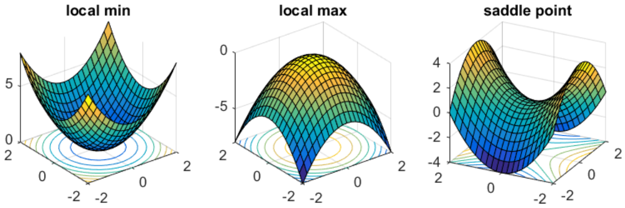
Pros of policy gradient [Silver ′14]:

- Better convergence properties (How much better?)
- Effective in high-dimensional or continuous action spaces
- Can learn stochastic policies

Cons of policy gradient [Silver ′14]:

- Typically converge to a local rather than global optimum (Really?)

⇒ First-order algorithms are not guaranteed to find local optima



Context

Pros of policy gradient [Silver ′14]:

Better convergence properties (How much better?)

Effective in high-dimensional or continuous action spaces

Can learn stochastic policies

Cons of policy gradient [Silver ′14]:

Typically converge to a local rather than global optimum (Really?)

Contribution: global convergence of policy gradient methods

$\Rightarrow$ for discounted infinite-horizon setting w/ iteration complexity

$\Rightarrow$ conditions for converging to approximate local extrema

Contrast w/ asymptotics via ODEs [Kushner & Yin ′76; Borkar ′08]

$\Rightarrow$ Correct claims of attaining local extrema via nonconvex opt.

Policy Gradient Theorem

Policy gradient formula [Sutton ′00]

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{(s,a)\sim\rho_\theta(\cdot,\cdot)}\left[\nabla \log \pi_\theta(a \,|\, s) \cdot Q_{\pi_\theta}(s,a)\right].$$

$\Rightarrow \rho_\theta(s,a) \Rightarrow$ ergodic dist. of Markov chain for fixed policy:

$$\rho_\theta(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p(s_t = s \,|\, s_0, \pi_\theta) \cdot \pi_\theta(a \,|\, s).$$

Policy Gradient Theorem

Policy gradient formula [Sutton ′00]

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{(s,a)\sim\rho_\theta(\cdot,\cdot)}\left[\nabla \log \pi_\theta(a \,|\, s) \cdot Q_{\pi_\theta}(s,a)\right].$$

$\Rightarrow \rho_\theta(s,a) \Rightarrow$ ergodic dist. of Markov chain for fixed policy:

$$\rho_\theta(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t p(s_t = s \,|\, s_0, \pi_\theta) \cdot \pi_\theta(a \,|\, s).$$

Stochastic gradient ascent (SGA): $\theta_{k+1} = \theta_k + \alpha_k \hat{\nabla} J(\theta_k)$.

Policy Gradient Theorem

Policy gradient formula [Sutton '00]

$$\nabla J(\theta) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{(s,a)\sim\rho_\theta(\cdot,\cdot)}\left[\nabla \log \pi_\theta(a \,|\, s) \cdot Q_{\pi_\theta}(s,a)\right].$$

$\Rightarrow \rho_\theta(s,a)$ $\Rightarrow$ ergodic dist. of Markov chain for fixed policy:

$$\rho_\theta(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p(s_t = s \,|\, s_0, \pi_\theta)\cdot \pi_\theta(a \,|\, s).$$

Stochastic gradient ascent (SGA): $\theta_{k+1} = \theta_k + \alpha_k \hat{\nabla} J(\theta_k)$.

Unbiasedly sampling $\hat{\nabla} J(\theta)$ is challenging, since this requires

$\Rightarrow$ $\hat{Q}_{\pi_\theta}(s,a)$ unbiasedly estimate $Q_{\pi_\theta}(s,a)$

$\Rightarrow$ (s,a) drawn from $\rho_\theta(\cdot,\cdot)$

Random-horizon Policy Gradient (RPG)

Unbiasedly estimate $Q_{\pi_\theta}(s,a)$ [Paternain 2018]:

$\Rightarrow$ Draw $T' \sim \text{Geom}(1-\gamma^{1/2})$, i.e., $P(T'=t) = (1-\gamma^{1/2})\gamma^{t/2}$

$\Rightarrow$ Rollout a trajectory $(s_0, a_0, s_1, \cdots, s_{T'}, a_{T'})$

$$\hat{Q}_{\pi_\theta}(s,a) = \sum_{t=0}^{T'} \gamma^{t/2} \cdot R(s_t, a_t) \,\Big|\, s_0 = s, a_0 = a$$

Random-horizon Policy Gradient (RPG)

Unbiasedly estimate $Q_{\pi_\theta}(s,a)$ [Paternain 2018]:

$\Rightarrow$ Draw $T' \sim \mathrm{Geom}(1-\gamma^{1/2})$, i.e., $P(T'=t) = (1-\gamma^{1/2})\gamma^{t/2}$

$\Rightarrow$ Rollout a trajectory $(s_0, a_0, s_1, \cdots, s_{T'}, a_{T'})$

$$\hat{Q}_{\pi_\theta}(s,a) = \sum_{t=0}^{T'} \gamma^{t/2} \cdot R(s_t, a_t) \,\Big|\, s_0 = s, a_0 = a$$

$\Rightarrow$ Benefit of $\gamma^{1/2}$: almost sure (a.s.) boundedness of $\hat{Q}_{\pi_\theta}(s,a)$

Random-horizon Policy Gradient (RPG)

Unbiasedly estimate $Q_{\pi_\theta}(s,a)$ [Paternain 2018]:

$\Rightarrow$ Draw $T' \sim \text{Geom}(1-\gamma^{1/2})$, i.e., $P(T'=t) = (1-\gamma^{1/2})\gamma^{t/2}$

$\Rightarrow$ Rollout a trajectory $(s_0, a_0, s_1, \cdots, s_{T'}, a_{T'})$

$$\hat{Q}_{\pi_\theta}(s,a) = \sum_{t=0}^{T'} \gamma^{t/2} \cdot R(s_t, a_t) \,\Big|\, s_0 = s, a_0 = a$$

$\Rightarrow$ Benefit of $\gamma^{1/2}$: almost sure (a.s.) boundedness of $\hat{Q}_{\pi_\theta}(s,a)$

Draw (s,a) from $\rho_\theta(\cdot,\cdot)$:

$\Rightarrow$ Draw $T \sim \text{Geom}(1-\gamma)$

$\Rightarrow$ Rollout a trajectory $(s_0, a_0, s_1, \cdots, s_T, a_T)$

$\Rightarrow$ Evaluate the gradient at (s_T, a_T)

$$\hat{\nabla} J(\theta) = \frac{1}{1-\gamma} \cdot \hat{Q}_{\pi_\theta}(s_T, a_T) \cdot \nabla \log[\pi_\theta(a_T \,|\, s_T)]$$

Random-horizon Policy Gradient (RPG)

Unbiasedly estimate $Q_{\pi_\theta}(s, a)$ [Paternain 2018]:

$\Rightarrow$ Draw $T' \sim \text{Geom}(1 - \gamma^{1/2})$, i.e., $P(T' = t) = (1 - \gamma^{1/2})\gamma^{t/2}$

$\Rightarrow$ Rollout a trajectory $(s_0, a_0, s_1, \cdots, s_{T'}, a_{T'})$

$$\hat{Q}_{\pi_\theta}(s, a) = \sum_{t=0}^{T'} \gamma^{t/2} \cdot R(s_t, a_t) \,\Big|\, s_0 = s, a_0 = a$$

$\Rightarrow$ Benefit of $\gamma^{1/2}$: almost sure (a.s.) boundedness of $\hat{Q}_{\pi_\theta}(s, a)$

Draw (s, a) from $\rho_\theta(\cdot, \cdot)$:

$\Rightarrow$ Draw $T \sim \text{Geom}(1 - \gamma)$

$\Rightarrow$ Rollout a trajectory $(s_0, a_0, s_1, \cdots, s_T, a_T)$

$\Rightarrow$ Evaluate the gradient at (s_T, a_T)

$$\hat{\nabla} J(\theta) = \frac{1}{1 - \gamma} \cdot \hat{Q}_{\pi_\theta}(s_T, a_T) \cdot \nabla \log[\pi_\theta(a_T \,|\, s_T)]$$

Random-horizon Policy Gradient (RPG) update:

$$\theta_{k+1} = \theta_k + \alpha_k \hat{\nabla} J(\theta_k)$$

Convergence Guarantee

Asymptotic convergence to stationary points:

Theorem (Convergence with Diminishing Stepsize)

Let $\{\theta_k\}_{k\geq 0}$ be the sequence of parameters of the policy π_{θ_k} given by RPG. If the stepsize $\{\alpha_k\}$ satisfies

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty,$$

then we have

$$\lim_{k\to\infty} \|\nabla J(\theta_k)\| = 0, \quad a.s.$$

$\Rightarrow$ Recover the result obtained by ODE method (Borkar & Meyn)

Convergence Guarantee

Convergence rate with diminishing stepsize

Theorem (Rate with Diminishing Stepsize)

Let $\{\theta_k\}_{k\geq 0}$ *be the sequence of parameters of the policy* π_{θ_k} *given by RPG. Let the stepsize be* $\alpha_k = k^{-a}$ *where* $a \in (0, 1)$. *Let*

$$K_\epsilon = \min\left\{k : \inf_{0\leq m\leq k} \mathbb{E}[\|\nabla J(\theta_m)\|^2] \leq \epsilon\right\} \leq \mathcal{O}(\epsilon^{-\frac{1}{2}})$$

$\Rightarrow$ Recover the $O(1/\sqrt{k})$ optimal rate of SGA for nonconvex opt.

Convergence Guarantee

Convergence with constant stepsize

Corollary (Convergence with Constant Stepsize)

Let $\{\theta_k\}_{k\geq 0}$ be the sequence of parameters of the policy π_{θ_k} given by RPG. Let the stepsize be $\alpha_k = \alpha > 0$. Then, there exists some constant $C > 0$ such that

$$\frac{1}{k}\sum_{m=1}^{k}\mathbb{E}[\|\nabla J(\theta_m)\|^2] \leq O\bigg(\frac{1}{k\alpha} + C\cdot\alpha\bigg).$$

$\Rightarrow$ Recover the conv. of SGA to the neighborhood of stationary points

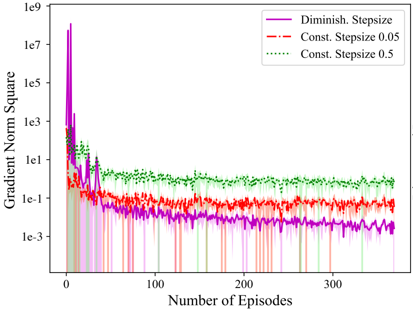
$\Rightarrow$ Trade-off between the conv. speed and the accuracy by choosing α

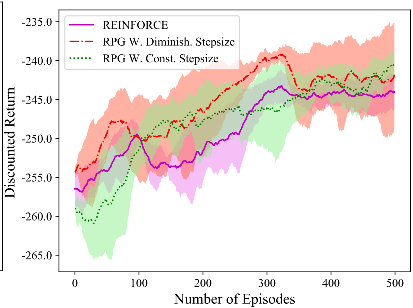
Pendulum Experiments

Compare with REINFORCE [Williams ′92]

⇒ fixed Q function horizon estimate

Each curve 30 times with mean and ±1.0 standard deviation





Additional Assumptions

Can we do better? Link R & π_θ to 2nd-order structure of value func.

Assumption

Positive/negative reward: $|R(s,a)| \in [L_R, U_R]$ uniformly with $L_R > 0$.
Fisher information matrix induced by $\pi_\theta(\cdot \mid s)$ is positive-definite

$$G(\theta) := \int_{\mathcal{S}\times\mathcal{A}} \rho_\theta(s,a) \cdot \nabla \log \pi_\theta(a \mid s) \cdot [\nabla \log \pi_\theta(a \mid s)]^\top dads \succeq L_I \cdot \boldsymbol{I}.$$

Smoothness: there exist $\rho_\Theta > 0$ and $C_\Theta < \infty$ s.t. for any $(s,a) \in \mathcal{S} \times \mathcal{A}$

$$\left\|\nabla^2 \log \pi_{\theta^1}(a \mid s) - \nabla^2 \log \pi_{\theta^2}(a \mid s)\right\| \leq \rho_\Theta \cdot \|\theta^1 - \theta^2\|, \text{for all } \theta^1, \theta^2,$$

$$\left\|\nabla^2 \log \pi_\theta(a \mid s)\right\| \leq C_\Theta, \text{for all } \theta.$$

Can be easily satisfied in practice.

$\Rightarrow$ motivates reward offset via nonconvex opt $\Rightarrow$ common in practice

Modified RPG Algorithm

Algorithm 1 **MRPG:** Modified Random-horizon Policy Gradient Algorithm

Input: s_0, θ_0, and the gradient type $\diamondsuit$, initialize $k \leftarrow 0$, return set $\hat{\Theta}^* \leftarrow \emptyset$.

Repeat:

Draw T_{k+1} from $\text{Geom}(1-\gamma)$, and draw $a_0 \sim \pi_{\theta_k}(\cdot \mid s_0)$.

for all $t = 0, \cdots, T_{k+1} - 1$ **do**

Simulate $s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$ and $a_{t+1} \sim \pi_{\theta_k}(\cdot \mid s_{t+1})$.

end for

Calculate the stochastic gradient $g_k \leftarrow$ **EvalPG**$(s_{T_{k+1}}, a_{T_{k+1}}, \theta_k, \diamondsuit)$.

if $(k \bmod k_{\text{thre}}) = 0$ **then**

$$\hat{\Theta}^* \leftarrow \hat{\Theta}^* \cup \{\theta_k\}, \qquad \theta_{k+1} \leftarrow \theta_k + \beta \cdot g_k$$

else

$$\theta_{k+1} \leftarrow \theta_k + \alpha \cdot g_k$$

end if

Update the iteration counter $k = k + 1$.

Until Convergence

return θ uniformly at random from the set $\hat{\Theta}^*$.

Improved Convergence Guarantee

Definition (Second-order Stationary Point)

A point θ is an ϵ_g, ϵ_h-second order stationary point with $\epsilon_g, \epsilon_h > 0$, if

$$\|\nabla J(\theta)\| \leq \epsilon_g, \quad \nabla^2 J(\theta) \preceq \epsilon_h \cdot \boldsymbol{I}.$$

Approximate local optima if no degenerate saddle exists

Improved Convergence Guarantee

Definition (Second-order Stationary Point)

A point θ is an ϵ_g, ϵ_h-second order stationary point with $\epsilon_g, \epsilon_h > 0$, if

$$\|\nabla J(\theta)\| \leq \epsilon_g, \quad \nabla^2 J(\theta) \preceq \epsilon_h \cdot \boldsymbol{I}.$$

Approximate local optima if no degenerate saddle exists

Theorem (Improved Convergence)

Let $\{\theta_k\}_{k\geq 0}$ be the sequence of parameters of the policy π_{θ_k} given by the MRPG updates, with certain parameters chosen, then θ_k converges to an $(\epsilon, \sqrt{\epsilon})$-second order stationary point w/ prob. $(1-\delta)$ after

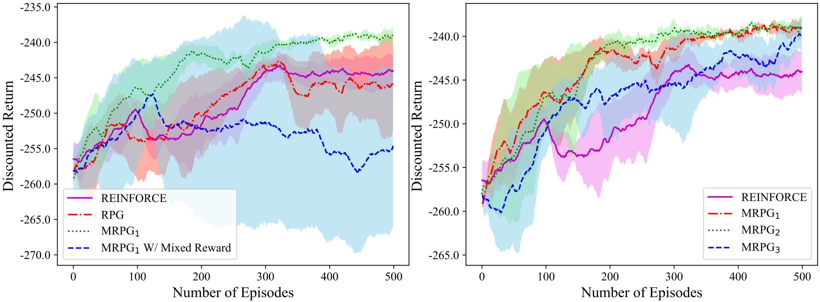
$$\mathcal{O}\left(\left(\frac{\rho^{3/2} L \epsilon^{-9}}{\delta \eta}\right) \log\left(\frac{\ell_g L}{\epsilon \eta \rho}\right)\right),$$

steps. If no degenerate saddle exists, attain locally optimal policy.

Pendulum Experiments

Compare with REINFORCE [Williams ′92]

Each curve 30 times with mean and ± 1.0 standard deviation



Mixed reward setting: adding a constant 10.0

Conclusion

Policy gradient method $\Rightarrow$ foundation of many RL methods

$\Rightarrow$ global convergence and limiting properties not well-understood

$\Rightarrow$ in infinite horizon settings

We derive iteration complexity from nonconvex opt perspective

$\Rightarrow$ of a new version that uses random rollout horizons for Q function

$\Rightarrow$ establish conditions for attaining approximate local extrema

Experimentally observe these properties of policy search on pendulum

$\Rightarrow$ solid foundation to derive accelerated & variance-reduced methods