# Policy Search in Reinforcement Learning: Advances Through Non-Convex Optimization

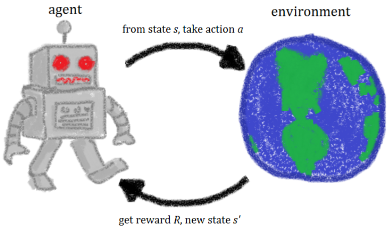
Kaiqing Zhang[⋆] **Alec Koppel**[†] Hao Zhu[‡] Tamer Başar[⋆]
[⋆]UIUC [†]ARL [‡]UT Austin

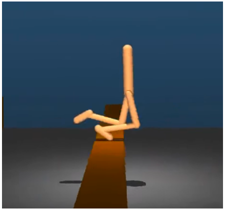
Estimation and Learning in Stochastic Systems
IEEE Conference on Information Science and Systems

Mar. 21, 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]



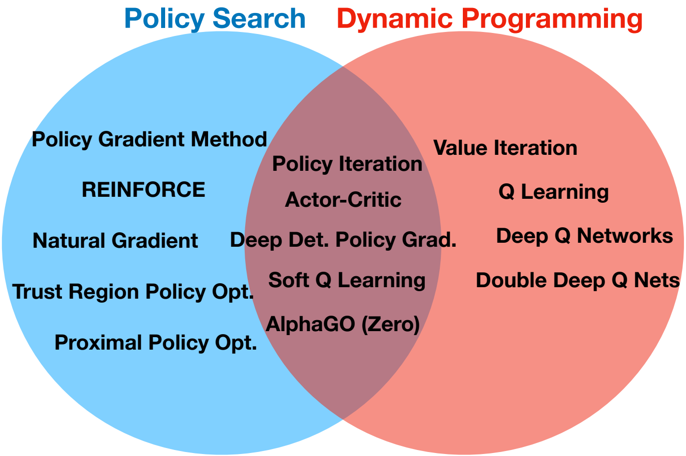
# Problem Formulation

- 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 $\mathbb{P}(s' \,|\, s, a) : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$
  - $\Rightarrow$ 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., $a_t \sim \pi(\cdot \,|\, s_t)$
- Infinite-horizon setting value function:

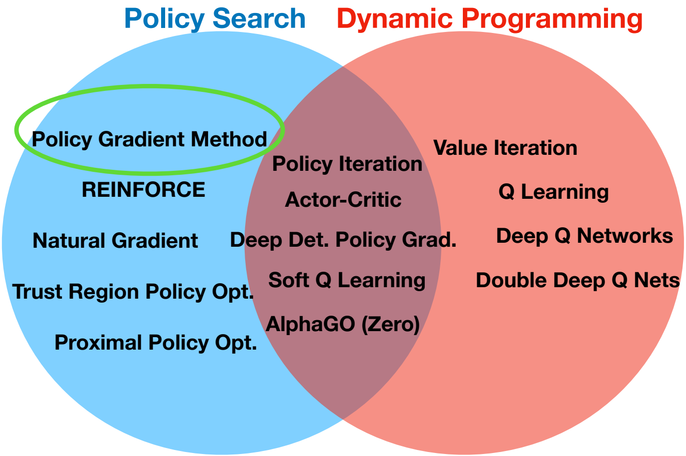
$$V(s) = \mathbb{E}\left( \sum_{t=0}^{\infty} \gamma^t \cdot R(s_t, a_t) \,\middle|\, s_0 = s \right),$$

- Goal: find $\{a_t = \pi(s_t)\}$ to maximize $V_\pi(s) := \mathbb{E}[V(s) \,|\, a \sim \pi(s)]$
- $\max_{\pi \in \Pi} V_\pi(s)$ where $\Pi$ is some family of distributions
  - $\Rightarrow$ E.g., Gaussian $\pi = \pi_\theta$ w/ $\theta \in \mathbb{R}^d$ $\Rightarrow \pi_\theta(\cdot \,|\, s) = \mathcal{N}(\phi(s)^\top \theta, \sigma^2)$
  - $\Rightarrow$ Define action-state value (Q) function $Q_\pi(s, a) = \mathbb{E}[V_\pi(s) \,|\, a_0 = a]$

# Literature Landscape



# Literature Landscape



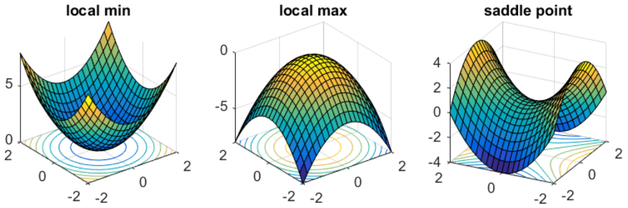
# Context

- 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?)

$\Rightarrow$ 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**

  ⇒ for discounted infinite-horizon setting w/ iteration complexity

  ⇒ conditions for converging to approximate local extrema
- Contrast w/ asymptotics via ODEs [Kushner & Yin ′76; Borkar ′08]

  ⇒ Correct claims of attaining local extrema via nonconvex opt.

# Policy Gradient Theorem

- Policy gradient formula [Sutton '00]

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

$\Rightarrow \rho_\theta(s,a) \Rightarrow$ 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]

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

$\Rightarrow \rho_\theta(s,a) \ \Rightarrow$ 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]

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

$\Rightarrow \rho_\theta(s,a) \Rightarrow$ 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 $T' \sim$ 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 $\pi_{\theta_k}$ 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*

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

$\Rightarrow$ Recover the result obtained by ODE method

# 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 $\pi_{\theta_k}$ given by Algorithm 3. 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* $\pi_{\theta_k}$ *given by Algorithm 3. Let the stepsize be* $\alpha_k = \alpha > 0$. *Then, there exists some constant* $C > 0$ *such that*

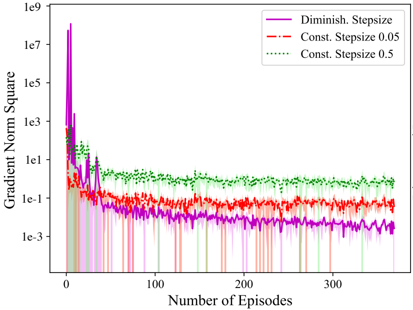
$$\frac{1}{k}\sum_{m=1}^{k}\mathbb{E}[\|\nabla J(\theta_m)\|^2] \leq O\Big(\frac{1}{k\alpha} + C\cdot\alpha\Big).$$

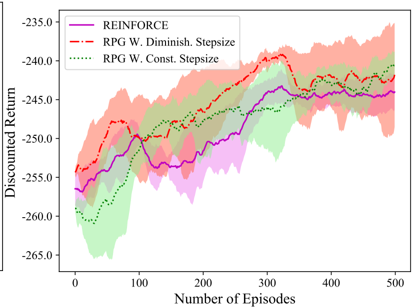
$\Rightarrow$ Recover the conv. of SGA to the neighborhood of stationary points

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

# Pendulum Experiments

- Compare with REINFORCE [Williams ′92]
  $\Rightarrow$ fixed Q function horizon estimate
- Each curve 30 times with mean and $\pm 1.0$ standard deviation



# Additional Assumptions

- Can we do better? Link $R$ & $\pi_\theta$ 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 $\pi_\theta(\cdot \,|\, s)$ is positive-definite

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

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

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

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

- 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:** $s_0, \theta_0$, 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** $\theta$ uniformly at random from the set $\hat{\Theta}^*$.

# Improved Convergence Guarantee

Definition (Second-order Stationary Point)

A point $\theta$ is an $\epsilon_g, \epsilon_h$-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 $\theta$ is an $\epsilon_g, \epsilon_h$-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)**

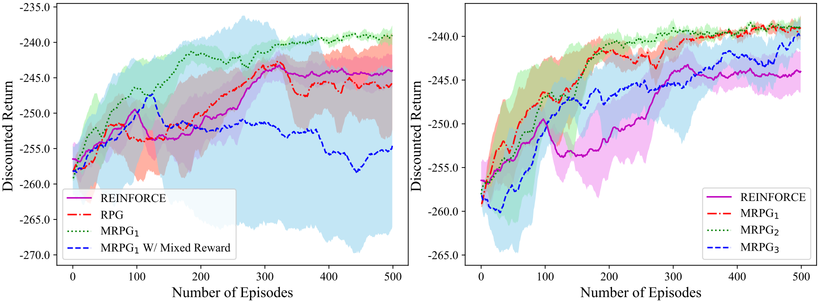
*Let $\{\theta_k\}_{k\geq 0}$ be the sequence of parameters of the policy $\pi_{\theta_k}$ given by the MRPG updates, i.e., Algorithm 1, with certain parameters chosen, then $\theta_k$ converges to an $(\epsilon, \sqrt{\epsilon})$-second order stationary point w/ prob. $(1-\delta)$ after*

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

*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
  $\Rightarrow$ its global convergence and limiting properties not well-understood
- We derive iteration complexity from nonconvex opt perspective
  $\Rightarrow$ of a new version that uses random rollout horizons for $Q$ function
  $\Rightarrow$ establish conditions under to attain approximate local extrema
- Experimentally observe these properties of policy search on pendulum
  $\Rightarrow$ solid foundation to derive accelerated $\&$ variance-reduced methods

# Problem Formulation

- Objective: Find the policy that maximizes the value given $s_0$

$$\max_{\pi \in \Pi} V_\pi(s_0).$$

# Problem Formulation

- Objective: Find the policy that maximizes the value given $s_0$

$$\max_{\pi \in \Pi} V_\pi(s_0).$$

- Parameterized policy $\pi = \pi_\theta$ with $\theta \in \mathbb{R}^d$, e.g., Gaussian policy

$$\pi_\theta(\cdot \,|\, s) = \mathcal{N}(\phi(s)^\top \theta, \sigma^2)$$

# Problem Formulation

- Objective: Find the policy that maximizes the value given $s_0$

$$\max_{\pi \in \Pi} V_\pi(s_0).$$

- Parameterized policy $\pi = \pi_\theta$ with $\theta \in \mathbb{R}^d$, e.g., Gaussian policy

$$\pi_\theta(\cdot \,|\, s) = \mathcal{N}(\phi(s)^\top \theta, \sigma^2)$$

- A nonconvex optimization problem

$$\boxed{\max_{\theta} J(\theta) := V_{\pi_\theta}(s_0).}$$

# Problem Formulation

- Objective: Find the policy that maximizes the value given $s_0$

$$\max_{\pi \in \Pi} V_\pi(s_0).$$

- Regularity conditions of the reward $R$ and $\pi_\theta$

**Assumption**

- Boundedness: $|R(s,a)| \in [0, U_R]$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.
- Smoothness: $\pi_\theta$ is differentiable with respect to $\theta$, and $\nabla \log \pi_\theta(a \,|\, s)$ is $L$-Lipschitz and has bounded norm, i.e., for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\|\nabla \log \pi_{\theta^1}(a \,|\, s) - \nabla \log \pi_{\theta^2}(a \,|\, s)\| \leq L \cdot \|\theta^1 - \theta^2\|, \text{for all } \theta^1, \theta^2,$$

$$\|\nabla \log \pi_\theta(a \,|\, s)\| \leq B_\Theta, \text{for all } \theta.$$

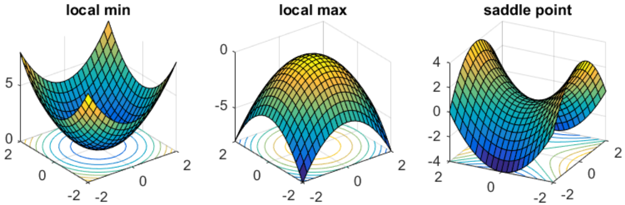
# Related Work

- Policy gradient theorems [Sutton ′00; Silver et al. ′14]
- Classical algorithms: REINFORCE [Williams ′92], Natural policy gradient [Kakade ′02], deterministic policy gradient [Silver et al. ′14]
- Tremendous empirical works, especially with deep neural nets [Lillicrap et al. ′15; Mnih et al. ′16]
- Actor-critic algorithms [Konda et al. ′00; Peters et al. ′09; Mnih et al. ′16] to reduce the variance, two-timescale algorithms
- Recently, Stochastic Variance-Reduced Policy Gradient [Papini et al. ′18], from an optimization perspective

- However, none of them established global convergence under a discounted infinite-horizon setting, with iteration complexity

# Challenges

- Unbiased estimate of policy gradient is elusive to obtain
  - $\Rightarrow$ Monte-Carlo for finite-horizon, e.g., REINFORCE, creates bias
  - $\Rightarrow$ Online actor-critic has both bias and correlated noise from the critic
- Mathematic tool is very general, but the results are limited
  - $\Rightarrow$ Stochastic approx. & ODE method [Kushner & Yin $'$76; Borkar $'$08]
  - $\Rightarrow$ Mostly asymptotic convergence only, i.e., when $t \to \infty$
- Understanding gap from a nonconvex optimization perspective
  - $\Rightarrow$ First-order algorithms are not guaranteed to find local optima



**Algorithm 2 EstQ:** Unbiasedly Estimating Q-function

**Input:** $s, a$, and $\theta$. Initialize $\hat{Q} \leftarrow 0$, $s_0 \leftarrow s$, and $a_0 \leftarrow a_0$.
Draw $T$ from the geometric distribution $\text{Geom}(1-\gamma^{1/2})$, i.e., $P(T=t) = (1-\gamma^{1/2})\gamma^{t/2}$.
**for all** $t = 0, \cdots, T-1$ **do**
  Collect and add the instantaneous reward $R(s_t, a_t)$ to $\hat{Q}$, $\hat{Q} \leftarrow \hat{Q} + \gamma^{t/2} \cdot R(s_t, a_t)$.
  Simulate the next state $s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$ and action $a_{t+1} \sim \pi(\cdot \mid s_{t+1})$.
**end for**
Collect $R(s_T, a_T)$ by $\hat{Q} \leftarrow \hat{Q} + \gamma^{T/2} \cdot R(s_T, a_T)$.
**return** $\hat{Q}$.

**Algorithm 3** **RPG:** Random-horizon Policy Gradient Algorithm

**Input:** $s_0$ and $\theta_0$, initialize $k \leftarrow 0$.
**Repeat:**
Draw $T_{k+1}$ from the geometric distribution Geom$(1-\gamma)$.
Draw $a_0 \sim \pi_{\theta_k}(\cdot \mid s_0)$
**for all** $t = 0, \cdots, T_{k+1} - 1$ **do**
Simulate the next state $s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$ and action $a_{t+1} \sim \pi_{\theta_k}(\cdot \mid s_{t+1})$.
**end for**
Obtain an estimate $\hat{Q}_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}}) \leftarrow \textbf{EstQ}(s_{T_{k+1}}, a_{T_{k+1}}, \theta_k)$.
Perform stochastic policy gradient

$$\theta_{k+1} \leftarrow \theta_k + \frac{\alpha_k}{1-\gamma} \cdot \hat{Q}_{\pi_{\theta_k}}(s_{T_{k+1}}, a_{T_{k+1}}) \cdot \nabla \log[\pi_{\theta_k}(a_{T_{k+1}} \mid s_{T_{k+1}})]$$

Update the iteration counter $k \leftarrow k+1$.
**Until Convergence**

# Theoretical Results

### Theorem (Unbiasedness)

*For any $\theta$ and $(s, a)$, $\hat{Q}_{\pi_\theta}(s, a)$ and $\hat{\nabla} J(\theta)$ are unbiased estimates of $Q_{\pi_\theta}(s, a)$ and $\nabla J(\theta)$, respectively, i.e.,*

$$\mathbb{E}[\hat{Q}_{\pi_\theta}(s, a)] = Q_{\pi_\theta}(s, a), \quad \mathbb{E}[\hat{\nabla} J(\theta)] = \nabla J(\theta).$$

# Theoretical Results

### Theorem (Unbiasedness)

*For any $\theta$ and $(s,a)$, $\hat{Q}_{\pi_\theta}(s,a)$ and $\hat{\nabla} J(\theta)$ are unbiased estimates of $Q_{\pi_\theta}(s,a)$ and $\nabla J(\theta)$, respectively, i.e.,*

$$\mathbb{E}[\hat{Q}_{\pi_\theta}(s,a)] = Q_{\pi_\theta}(s,a), \quad \mathbb{E}[\hat{\nabla} J(\theta)] = \nabla J(\theta).$$

- The gradient and its estimate have other nice properties

### Lemma (Properties of RPG)

- $\nabla J(\theta)$ *is bounded:* $\|\nabla J(\theta)\| \leq B_\Theta \cdot U_R/(1-\gamma)^2$ *and is* $L_\Theta$*-Lipschitz:*

$$\|\nabla J(\theta^1) - \nabla J(\theta^2)\| \leq L_\Theta \cdot \|\theta^1 - \theta^2\|$$

  *for some* $L_\Theta(U_R, L, B_\Theta, \gamma)$.

- $\hat{\nabla} J(\theta)$ *is almost surely bounded:*

$$\|\hat{\nabla} J(\theta)\| \leq \frac{B_\Theta U_R}{(1-\gamma)(1-\gamma^{1/2})}.$$

# Nonconvex Perspective

- Can we do better? More than stationary points?

# Nonconvex Perspective

- Can we do better? More than stationary points?
- A fundamental question in nonconvex opt.: **Can local optima be achieved using (stochastic) first-order methods, e.g., SGD?**

# Nonconvex Perspective

- Can we do better? More than stationary points?
- A fundamental question in nonconvex opt.: **Can local optima be achieved using (stochastic) first-order methods, e.g., SGD?**

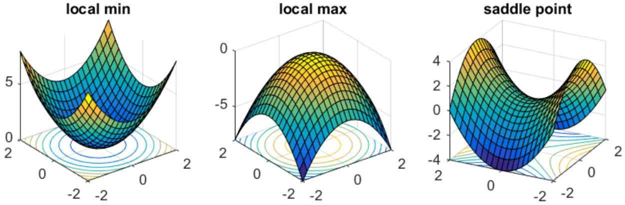
  $\Rightarrow$ **Yes!**

# Nonconvex Perspective

- Can we do better? More than stationary points?
- A fundamental question in nonconvex opt.: **Can local optima be achieved using (stochastic) first-order methods, e.g., SGD?**

  $\Rightarrow$ **Yes!**

  $\Rightarrow$ Key: if one can escape saddle points quickly

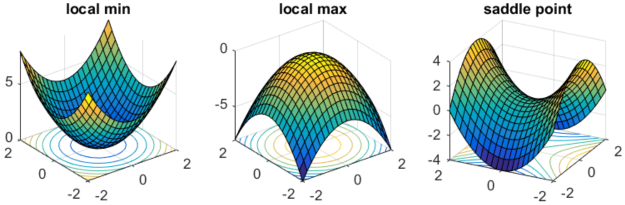


# Nonconvex Perspective

- Can we do better? More than stationary points?
- A fundamental question in nonconvex opt.: **Can local optima be achieved using (stochastic) first-order methods, e.g., SGD?**

  ⇒ **Yes!**

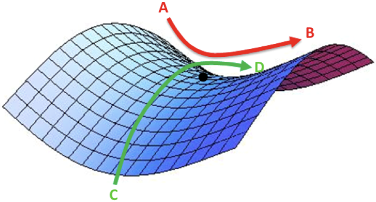
  ⇒ Key: if one can escape saddle points quickly




- Observed in many empirical results, e.g., in training neural nets [Bengio et al. ′14; Goodfellow et al. ′14]
- Recent theoretical advances [Ge et al. ′15; Lee et al. ′16; Jin et al. ′17]

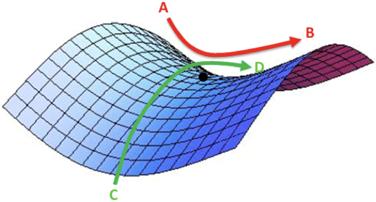
# Nonconvex Perspective

- Key idea:
  - $\Rightarrow$ The noise/perturbation of SGA is isotropic [Jin et al. $'15, '17$]



# Nonconvex Perspective

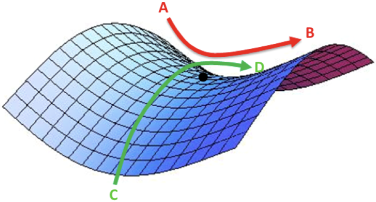
- Key idea:
  - ⇒ The noise/perturbation of SGA is isotropic [Jin et al. ′15, ′17]




- But the noise of SPG is not controlled in RL

# Nonconvex Perspective

- Key idea:
  - ⇒ The noise/perturbation of SGA is isotropic [Jin et al. ′15, ′17]




- But the noise of SPG is not controlled in RL
- Saddle-escaping without isotropic noise [Daneshmand et al. ′18]
- But need Correlated Negative Curvature (CNC) conditon

# Nonconvex Perspective

Assumption (CNC Condition (Daneshmand et al. $'18$))

Let $\boldsymbol{v}_\theta$ be the eigenvector corresponding to the maximum eigenvalue of the Hessian matrix $\mathcal{H}(\theta)$. The stochastic gradient $\hat{\nabla} J(\theta)$ satisfies the CNC condition, if the second moment of its projection along the direction $\boldsymbol{v}_\theta$ is uniformly bounded away from zero, i.e.,

$$\exists \eta > 0, \quad s.t., \quad \text{for all } \theta \in \Theta, \quad \mathbb{E}\big\{[\boldsymbol{v}_\theta^\top \hat{\nabla} J(\theta)]^2\big\} > \eta.$$

# Nonconvex Perspective

**Assumption (CNC Condition (Daneshmand et al. ′18))**

Let $\boldsymbol{v}_\theta$ be the eigenvector corresponding to the maximum eigenvalue of the Hessian matrix $\mathcal{H}(\theta)$. The stochastic gradient $\hat{\nabla} J(\theta)$ satisfies the CNC condition, if the second moment of its projection along the direction $\boldsymbol{v}_\theta$ is uniformly bounded away from zero, i.e.,

$$\exists \eta > 0, \quad s.t., \quad \text{for all } \theta \in \Theta, \quad \mathbb{E}\big\{[\boldsymbol{v}_\theta^\top \hat{\nabla} J(\theta)]^2\big\} > \eta.$$

- Does our SPG $\hat{\nabla} J(\theta)$ satisfy CNC condition?

# Nonconvex Perspective

Assumption (CNC Condition (Daneshmand et al. '18))

Let $\boldsymbol{v}_\theta$ be the eigenvector corresponding to the maximum eigenvalue of the Hessian matrix $\mathcal{H}(\theta)$. The stochastic gradient $\hat{\nabla} J(\theta)$ satisfies the CNC condition, if the second moment of its projection along the direction $\boldsymbol{v}_\theta$ is uniformly bounded away from zero, i.e.,

$$\exists \eta > 0, \quad s.t., \quad \text{for all } \theta \in \Theta, \quad \mathbb{E}\big\{[\boldsymbol{v}_\theta^\top \hat{\nabla} J(\theta)]^2\big\} > \eta.$$

- Does our SPG $\hat{\nabla} J(\theta)$ satisfy CNC condition?
- Yes! (under certain conditions)

# Properties of RPGs

- Strict positive/negative reward and Q-function amplify the variance

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

# Properties of RPGs

- Strict positive/negative reward and Q-function amplify the variance

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

- Propose SPG with baselines

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

$$\tilde{\nabla} J(\theta) = \frac{1}{1-\gamma} \cdot [R(s_T, a_T) + \gamma \hat{V}_{\pi_\theta}(s'_T) - \hat{V}_{\pi_\theta}(s_T)] \cdot \nabla \log[\pi_\theta(a_T \mid s_T)].$$

**Lemma**

*The stochastic policy gradients $\check{\nabla} J(\theta)$ and $\tilde{\nabla} J(\theta)$ are also unbiased estimate of $\nabla J(\theta)$, i.e., $\mathbb{E}[\check{\nabla} J(\theta)] = \mathbb{E}[\tilde{\nabla} J(\theta)] = \nabla J(\theta)$.*

# Properties of RPGs

- Strict positive/negative reward and Q-function amplify the variance

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

- Propose SPG with baselines

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

$$\tilde{\nabla} J(\theta) = \frac{1}{1-\gamma} \cdot [R(s_T, a_T) + \gamma \hat{V}_{\pi_\theta}(s_T') - \hat{V}_{\pi_\theta}(s_T)] \cdot \nabla \log[\pi_\theta(a_T \mid s_T)].$$

**Lemma**

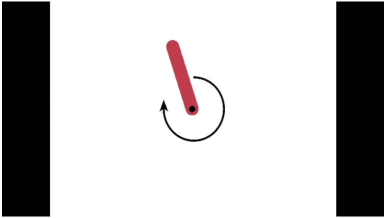
*The stochastic policy gradients $\check{\nabla} J(\theta)$ and $\tilde{\nabla} J(\theta)$ are also unbiased estimate of $\nabla J(\theta)$, i.e., $\mathbb{E}[\check{\nabla} J(\theta)] = \mathbb{E}[\tilde{\nabla} J(\theta)] = \nabla J(\theta)$.*

**Lemma**

*All the three stochastic policy gradients $\hat{\nabla} J(\theta)$, $\check{\nabla} J(\theta)$, and $\tilde{\nabla} J(\theta)$ satisfy the correlated negative curvature condition.*

# Simulations

- Environment: Pendulum in the OpenAI Gym [Brockman et al. '16]



- State: $s_t = (\cos(\theta_t), \sin(\theta_t), \dot{\theta}_t)^\top$; action $a_t \in [-20, 20]$ the joint effort
- Reward $R(s_t, a_t) \in [-17.1736044, -0.5]$:

$$R(s_t, a_t) := -(\theta^2 + 0.1 * \dot{\theta}^2 + 0.001 * a_t^2) - 0.5,$$

- Gaussian Policy: $\pi_\theta$ truncated over $[-20, 20]$ and parameterized by

$$\pi_\theta(\cdot \mid s) = \mathcal{N}(\mu_\theta(s), \sigma^2),$$

where $\mu_\theta(s)$ is a neural network with two hidden layers

# Beyond Vanilla Policy Gradient...

- Natural policy gradient [Kakade $'02$] performs superbly in practice

$$\bar{\nabla} J(\theta) := G(\theta)^{-1} \cdot \nabla J(\theta),$$

where $G(\theta)$ is the Fisher information matrix

# Beyond Vanilla Policy Gradient...

- Natural policy gradient [Kakade ′02] performs superbly in practice

$$\bar{\nabla} J(\theta) := G(\theta)^{-1} \cdot \nabla J(\theta),$$

  where $G(\theta)$ is the Fisher information matrix

- Stochastic Quasi-Newton method from an optimization perspective

# Beyond Vanilla Policy Gradient...

- Natural policy gradient [Kakade ′02] performs superbly in practice

$$\bar{\nabla} J(\theta) := G(\theta)^{-1} \cdot \nabla J(\theta),$$

where $G(\theta)$ is the Fisher information matrix

- Stochastic Quasi-Newton method from an optimization perspective

**Theorem (Global Convergence of Natural PG (Informal))**

*Let $\{\theta_k\}_{k\geq 0}$ be the sequence of parameters of the policy $\pi_{\theta_k}$ given by the Natural PG, then $\theta_k$ preserves the global convergence property of SPG to first-order stationary points.*

# Beyond Vanilla Policy Gradient...

- Natural policy gradient [Kakade '02] performs superbly in practice

$$\bar{\nabla} J(\theta) := G(\theta)^{-1} \cdot \nabla J(\theta),$$

where $G(\theta)$ is the Fisher information matrix

- Stochastic Quasi-Newton method from an optimization perspective

Theorem (Global Convergence of Natural PG (Informal))

*Let $\{\theta_k\}_{k\geq 0}$ be the sequence of parameters of the policy $\pi_{\theta_k}$ given by the Natural PG, then $\theta_k$ preserves the global convergence property of SPG to first-order stationary points.*

- What about convergence to second-order stationary points?
- Does $G(\theta)$ contain enough 2nd-order information to escape saddles?

# Concluding Remarks

- The global convergence property of PG methods is not well-understood

# Concluding Remarks

- The global convergence property of PG methods is not well-understood
- Propose an unbiased SPG estimate, facilitating a nonconvex perspective

# Concluding Remarks

- The global convergence property of PG methods is not well-understood
- Propose an unbiased SPG estimate, facilitating a nonconvex perspective
- Justify that simple conditions may indeed lead to local optimal policies, both theoretically and empirically

# Concluding Remarks

- The global convergence property of PG methods is not well-understood
- Propose an unbiased SPG estimate, facilitating a nonconvex perspective
- Justify that simple conditions may indeed lead to local optimal policies, both theoretically and empirically
- Summary & Future Work

| | 1st-order Stationary Points | 2nd-order Stationary Points |
|---|---|---|
| Vanilla SPG | ✓ Re-discover the asymptotic a.s. conv.; Establish conv. rate | ✓ Prove CNC condition; Establish conv. rate |
| Natural PG | ✓ Establish both asymptotic a.s. conv. and conv. rate | ? |

# Concluding Remarks

- The global convergence property of PG methods is not well-understood
- Propose an unbiased SPG estimate, facilitating a nonconvex perspective
- Justify that simple conditions may indeed lead to local optimal policies, both theoretically and empirically
- Summary & Future Work

| | 1st-order Stationary Points | 2nd-order Stationary Points |
|---|---|---|
| Vanilla SPG | ✓ Re-discover the asymptotic a.s. conv.; Establish conv. rate | ✓ Prove CNC condition; Establish conv. rate |
| Natural PG | ✓ Establish both asymptotic a.s. conv. and conv. rate | ? |

Thank You!