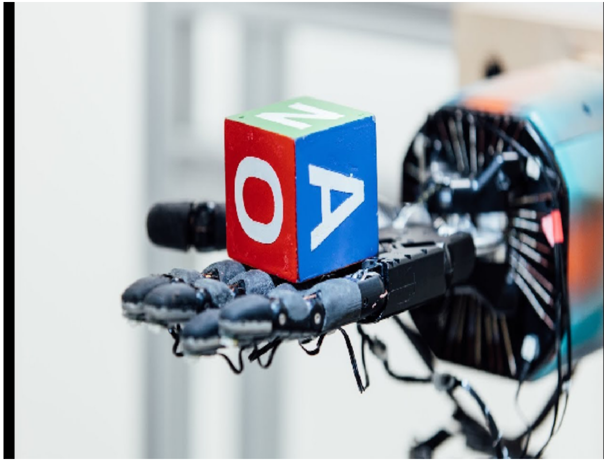
Policy Gradient

Policy Optimization

[AlphaZero, Silver et.al, 17]

[OpenAI Five, 18]


[OpenAI,19]

Recap of CPI

Recall CPI:

At Iteration t:

1. Greedy Policy Selector:

Recall we use regression

$$\pi' \in \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

2. If $\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi^t}} [A^{\pi^t}(s, \pi(s))] \leq \varepsilon$

   Return π^t

3. Incremental Update:

$$\pi^{t+1}(\cdot \mid s) = (1 - \alpha)\pi^t(\cdot \mid s) + \alpha \pi'(\cdot \mid s), \forall s$$

small number

Recap of CPI

1. Incremental update (Lemma 12.1 in AJKS)

$$\|d_\mu^{\pi^{t+1}}(\cdot) - d_\mu^{\pi^t}(\cdot)\|_1 \le \frac{2\gamma\alpha}{1-\gamma}$$

Recall CPI:

1. Greedy Policy Selector:

$$\pi' \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi(s))\right]$$

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s,\pi(s))] \le \varepsilon$

   Return π^t

3. Incremental Update:

$$\pi^{t+1}(\cdot\,|\,s) = (1-\alpha)\pi^t(\cdot\,|\,s) + \alpha\pi'(\cdot\,|\,s), \forall s$$

Recap of CPI

1. Incremental update (Lemma 12.1 in AJKS)

$$\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1 \leq \frac{2\gamma\alpha}{1-\gamma}$$

2. Before terminate, monotonic improvement (Thm 12.2 in AJKS):

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t} \geq \frac{\epsilon^2}{8\gamma}$$

(By setting step size α properly…)

Recall CPI:

1. Greedy Policy Selector:

$$\pi' \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi(s))\right]$$

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s,\pi(s))] \leq \varepsilon$

   Return π^t

3. Incremental Update:

$$\pi^{t+1}(\cdot\,|\,s) = (1-\alpha)\pi^t(\cdot\,|\,s) + \alpha\pi'(\cdot\,|\,s), \forall s$$

Recap of CPI

1. Incremental update (Lemma 12.1 in AJKS)

$$\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1 \leq \frac{2\gamma\alpha}{1-\gamma}$$

2. Before terminate, monotonic improvement (Thm 12.2 in AJKS):

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t} \geq \frac{\epsilon^2}{8\gamma}$$

(By setting step size α properly…)

Recall CPI:

1. Greedy Policy Selector:
$$\pi' \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi(s))\right]$$

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s,\pi(s))] \leq \varepsilon$

   Return π^t

3. Incremental Update:
$$\pi^{t+1}(\cdot\,|\,s) = (1-\alpha)\pi^t(\cdot\,|\,s) + \alpha\pi'(\cdot\,|\,s), \forall s$$

Local adv versus distribution change:

Recap of CPI

1. Incremental update (Lemma 12.1 in AJKS)

$$\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1 \leq \frac{2\gamma\alpha}{1-\gamma}$$

2. Before terminate, monotonic improvement (Thm 12.2 in AJKS):

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t} \geq \frac{\epsilon^2}{8\gamma}$$

(By setting step size α properly…)

Recall CPI:

1. Greedy Policy Selector:

$$\pi' \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi(s))\right]$$

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s,\pi(s))] \leq \varepsilon$

   Return π^t

3. Incremental Update:

$$\pi^{t+1}(\cdot\,|\,s) = (1-\alpha)\pi^t(\cdot\,|\,s) + \alpha\pi'(\cdot\,|\,s), \forall s$$

Local adv versus distribution change:

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t}$$

$$\geq \alpha\mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi'(s))\right] - \frac{\alpha}{1-\gamma}\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1$$

max-local-Adv

Recap of CPI

1. Incremental update (Lemma 12.1 in AJKS)

$$\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1 \leq \frac{2\gamma\alpha}{1-\gamma}$$

2. Before terminate, monotonic improvement (Thm 12.2 in AJKS):

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t} \geq \frac{\epsilon^2}{8\gamma}$$

(By setting step size α properly…)

Recall CPI:

1. Greedy Policy Selector:

$$\pi' \in \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi(s))\right]$$

Decision Tree

2. If $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi^t}}[A^{\pi^t}(s,\pi(s))] \leq \varepsilon$

Return π^t

3. Incremental Update:

$$\pi^{t+1}(\cdot\,|\,s) = (1-\alpha)\pi^t(\cdot\,|\,s) + \alpha\pi'(\cdot\,|\,s), \forall s$$

Local adv versus distribution change:

$$V_\mu^{\pi^{t+1}} - V_\mu^{\pi^t}$$

$$\geq \alpha\mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[A^{\pi^t}(s,\pi'(s))\right] - \frac{\alpha}{1-\gamma}\|d_\mu^{\pi^{t+1}} - d_\mu^{\pi^t}\|_1$$

$$\geq \alpha\epsilon - \frac{\gamma\alpha^2}{(1-\gamma)^2}$$

Recap: two definitions of MDPs

$$\mathcal{M} = \{P, r, \gamma, \mu, S, A\}$$ ← Infinite Discounted

where $s_0 \sim \mu$

Objective: $J(\pi) := \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 \sim \mu, s_{h+1} \sim P_{s_h, a_h}, a_h \sim \pi(\cdot \mid s_h)\right]$

Recap: two definitions of MDPs

$$\mathcal{M} = \{P, r, \gamma, \mu, S, A\}$$

where $s_0 \sim \mu$

Objective: $J(\pi) := \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 \sim \mu, s_{h+1} \sim P_{s_h, a_h}, a_h \sim \pi(\cdot \mid s_h)\right]$

Finite

$$\mathcal{M} = \{P, r, H, \mu, S, A\}$$

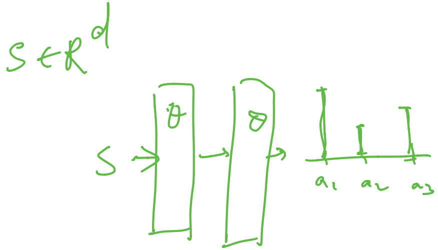
where $s_0 \sim \mu$

Objective: $J(\pi) := \mathbb{E}\left[\sum_{h=0}^{H-1} r(s_h, a_h) \,\middle|\, s_0 \sim \mu, s_{h+1} \sim P_{s_h, a_h}, a_h \sim \pi(\cdot \mid s_h)\right]$

Today: Policy Gradient Deriviation

Consider parameterized policy:

$$\pi_\theta(a \mid s) = \pi(a \mid s; \theta)$$



Today: Policy Gradient Derivation

Consider parameterized policy:

$$\pi_\theta(a \mid s) = \pi(a \mid s; \theta) \qquad J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{h=0}^{\infty} \gamma^h r_h\right]$$

Today: Policy Gradient Deriviation

Consider parameterized policy:

$$\pi_\theta(a \mid s) = \pi(a \mid s; \theta) \qquad J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{h=0}^{\infty} \gamma^h r_h\right]$$

$$\theta_{t+1} = \theta_t + \eta \boxed{\nabla_\theta J(\pi_\theta) \mid_{\theta=\theta_t}}$$

Gradient Ascent

Today: Policy Gradient Deriviation

Consider parameterized policy:

$$\pi_\theta(a \mid s) = \pi(a \mid s; \theta) \qquad J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{h=0}^{\infty} \gamma^h r_h\right]$$

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta J(\pi_\theta) \mid_{\theta=\theta_t}$$

Main question for today's lecture:
how to compute the gradient?

Outline for today

1. Recap on Gradient descent and stochastic gradient descent

2. Warm up: computing gradient using importance weighting

3. Policy Gradient formulations

Stochastic Gradient Descent

Given an objective function $J(\theta) : \mathbb{R}^d \mapsto \mathbb{R}$, (e.g., $J(\theta) = \mathbb{E}_{x,y}(f_\theta(x) - y)^2$)

SGD minimizes the above objective function as follows:

Stochastic Gradient Descent

Given an objective function $J(\theta) : \mathbb{R}^d \mapsto \mathbb{R}$, (e.g., $J(\theta) = \mathbb{E}_{x,y}(f_\theta(x) - y)^2$)

SGD minimizes the above objective function as follows:

Initialize θ_0, for t = 0, … :

Stochastic Gradient Descent

Given an objective function $J(\theta) : \mathbb{R}^d \mapsto \mathbb{R}$, (e.g., $J(\theta) = \mathbb{E}_{x,y}(f_\theta(x) - y)^2$)

SGD minimizes the above objective function as follows:

Initialize θ_0, for t = 0, … :

$$\theta_{t+1} = \theta_t - \eta g_t$$

↑ stoch - Gradient

Stochastic Gradient Descent

Given an objective function $J(\theta) : \mathbb{R}^d \mapsto \mathbb{R}$, (e.g., $J(\theta) = \mathbb{E}_{x,y}(f_\theta(x) - y)^2$)

SGD minimizes the above objective function as follows:

Initialize θ_0, for t = 0, … :

$$\theta_{t+1} = \theta_t - \eta g_t$$

unbiased estimate of $\nabla_\theta J(\theta_t)$

where $\mathbb{E}[g_t] = \nabla_\theta J(\theta_t)$

sample (x, y)

$$\nabla_\theta \left[(f_\theta(x) - y)^2 \right]$$

$$= 2(f_\theta(x) - y) \nabla_\theta f(x)$$

32, 64, 128

Sample $(x_i, y_i)_{i=1}^N$, i.i.d

$$\frac{1}{N} \sum_{i=1}^N 2(f_\theta(x_i) - y_i) \nabla_\theta f(x_i) \xrightarrow[N\to\infty]{} \nabla_\theta J(\theta)$$

Stochastic Gradient Descent

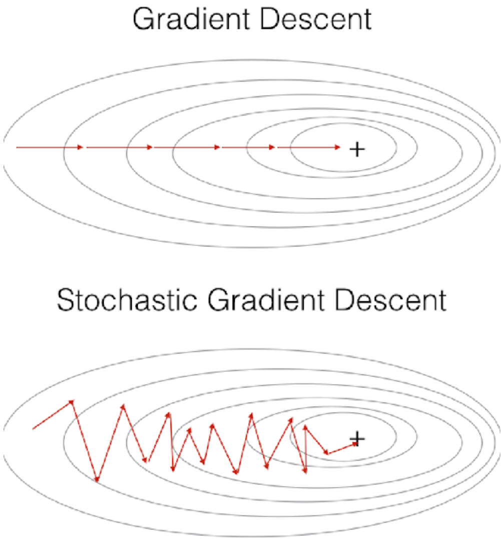
Given an objective function $J(\theta) : \mathbb{R}^d \mapsto \mathbb{R}$, (e.g., $J(\theta) = \mathbb{E}_{x,y}(f_\theta(x) - y)^2$)

SGD minimizes the above objective function as follows:

Initialize θ_0, for t = 0, … :

$$\theta_{t+1} = \theta_t - \eta g_t$$

where $\mathbb{E}[g_t] = \nabla_\theta J(\theta_t)$

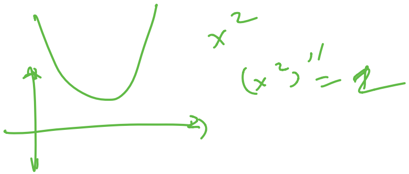


Convergence to Stationary Point

Consider a non-convex objective function $J(\theta)$,

Convergence to Stationary Point

Consider a non-convex objective function $J(\theta)$,



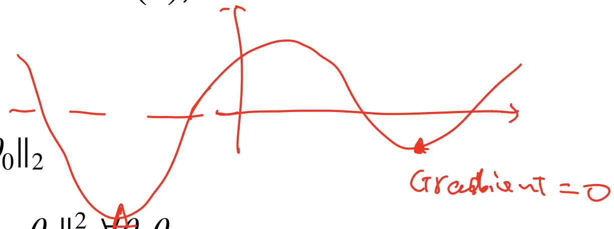
Def of β-smooth:

$$\|\nabla_\theta J(\theta) - \nabla_\theta J(\theta_0)\|_2 \leq \beta \|\theta - \theta_0\|_2$$

$$\left| J(\theta) - J(\theta_0) - \nabla_\theta J(\theta_0)^\top (\theta - \theta_0) \right| \leq \frac{\beta}{2} \|\theta - \theta_0\|_2^2, \forall \theta, \theta_0$$

Convergence to Stationary Point

Consider a non-convex objective function $J(\theta)$,



Def of β-smooth:

$$\|\nabla_\theta J(\theta) - \nabla_\theta J(\theta_0)\|_2 \le \beta\|\theta - \theta_0\|_2$$

$$\left| J(\theta) - J(\theta_0) - \nabla_\theta J(\theta_0)^\top(\theta - \theta_0) \right| \le \frac{\beta}{2}\|\theta - \theta_0\|_2^2, \forall \theta, \theta_0$$

[**Theorem**] If $J(\theta)$ is β-smooth and $\sup_\theta |J(\theta)| \le M$, and we run SGD: $\theta_{t+1} = \theta_t - \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t)$, $\quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \le \sigma^2$,

unbiased — second-moment of gradient

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \le O\left(\sqrt{\beta\sigma^2/T}\right)$$

# of SGD iteration

Proof of Convergence to Stationary Point (optional)

[**Theorem**] If $J(\theta)$ is β-smooth and $\sup_\theta |J(\theta)| \leq M$, and we run SGD: $\theta_{t+1} = \theta_t - \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \leq \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq O\left(\sqrt{\beta\sigma^2/T}\right)$$

Proof of Convergence to Stationary Point (optional)

[**Theorem**] If $J(\theta)$ is β-smooth and $\sup_\theta |J(\theta)| \le M$, and we run SGD: $\theta_{t+1} = \theta_t - \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \le \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \le O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left| J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t) \right| \le \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

Proof of Convergence to Stationary Point (optional)

[**Theorem**] If $J(\theta)$ is β-smooth and $\sup_\theta |J(\theta)| \leq M$, and we run SGD: $\theta_{t+1} = \theta_t - \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \leq \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \leq \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) + \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \leq \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

Proof of Convergence to Stationary Point (optional)

[**Theorem**] If $J(\theta)$ is β-smooth and $\sup_\theta |J(\theta)| \leq M$, and we run SGD: $\theta_{t+1} = \theta_t - \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \leq \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \leq \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) + \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \leq \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t) \leq -J(\theta_{t+1}) + J(\theta_t) + \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

Proof of Convergence to Stationary Point (optional)

[**Theorem**] If $J(\theta)$ is β-smooth and $\sup_\theta |J(\theta)| \le M$, and we run SGD: $\theta_{t+1} = \theta_t - \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \le \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \le O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \le \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) + \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \le \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t) \le -J(\theta_{t+1}) + J(\theta_t) + \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \mathbb{E}\left[\eta \nabla_\theta J(\theta_t)^\top \nabla_\theta J(\theta_t)\right] \le \mathbb{E}\left[J(\theta_t) - J(\theta_{t+1})\right] + \frac{\beta}{2}\eta^2\sigma^2$$

Proof of Convergence to Stationary Point (optional)

[**Theorem**] If $J(\theta)$ is β-smooth and $\sup_\theta |J(\theta)| \le M$, and we run SGD: $\theta_{t+1} = \theta_t - \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \le \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \le O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \le \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) + \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \le \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta \nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t) \le -J(\theta_{t+1}) + J(\theta_t) + \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \mathbb{E}\left[\eta \nabla_\theta J(\theta_t)^\top \nabla_\theta J(\theta_t)\right] \le \mathbb{E}\left[J(\theta_t) - J(\theta_{t+1})\right] + \frac{\beta}{2}\eta^2\sigma^2$$

$$\Rightarrow \eta\mathbb{E}\left[\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \le \sum_t \mathbb{E}\left[J(\theta_t) - J(\theta_{t+1})\right] + \frac{\beta T}{2}\eta^2\sigma^2$$

Proof of Convergence to Stationary Point (optional)

[Theorem] If $J(\theta)$ is β-smooth and $\sup_\theta |J(\theta)| \leq M$, and we run SGD: $\theta_{t+1} = \theta_t - \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \leq \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \leq \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) + \eta\nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \leq \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta\nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t) \leq -J(\theta_{t+1}) + J(\theta_t) + \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \mathbb{E}\left[\eta\nabla_\theta J(\theta_t)^\top \nabla_\theta J(\theta_t)\right] \leq \mathbb{E}\left[J(\theta_t) - J(\theta_{t+1})\right] + \frac{\beta}{2}\eta^2\sigma^2$$

$$\Rightarrow \eta\mathbb{E}\left[\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq \sum_t \mathbb{E}\left[J(\theta_t) - J(\theta_{t+1})\right] + \frac{\beta T}{2}\eta^2\sigma^2 \quad \Rightarrow \frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2 \leq \frac{1}{\eta T}M + \frac{\beta}{2}\eta\sigma^2$$

Proof of Convergence to Stationary Point (optional)

[Theorem] If $J(\theta)$ is β-smooth and $\sup_\theta |J(\theta)| \leq M$, and we run SGD: $\theta_{t+1} = \theta_t - \eta \widetilde{\nabla}_\theta J(\theta_t)$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t), \quad \mathbb{E}\left[\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \leq \sigma^2,$

then:

$$\mathbb{E}\left[\frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq O\left(\sqrt{\beta\sigma^2/T}\right)$$

$$\left|J(\theta_{t+1}) - J(\theta_t) - \nabla_\theta J(\theta_t)^\top(\theta_{t+1} - \theta_t)\right| \leq \frac{\beta}{2}\|\theta_{t+1} - \theta_t\|_2^2$$

$$\Rightarrow \left|J(\theta_{t+1}) - J(\theta_t) + \eta\nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t)\right| \leq \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

$$\Rightarrow \eta\nabla_\theta J(\theta_t)^\top \widetilde{\nabla}_\theta J(\theta_t) \leq -J(\theta_{t+1}) + J(\theta_t) + \frac{\beta}{2}\eta^2\|\widetilde{\nabla}_\theta J(\theta_t)\|_2^2$$

Set $\eta = \sqrt{M/(\beta\sigma^2 T)}$

$$\Rightarrow \mathbb{E}\left[\eta\nabla_\theta J(\theta_t)^\top \nabla_\theta J(\theta_t)\right] \leq \mathbb{E}\left[J(\theta_t) - J(\theta_{t+1})\right] + \frac{\beta}{2}\eta^2\sigma^2$$

$$\Rightarrow \eta\mathbb{E}\left[\sum_t \|\nabla_\theta J(\theta_t)\|_2^2\right] \leq \sum_t \mathbb{E}\left[J(\theta_t) - J(\theta_{t+1})\right] + \frac{\beta T}{2}\eta^2\sigma^2 \;\Rightarrow \frac{1}{T}\sum_t \|\nabla_\theta J(\theta_t)\|_2 \leq \frac{1}{\eta T}M + \frac{\beta}{2}\eta\sigma^2$$

Summary so far:

SGD is a simple algorithm that can find a locally optimal solution
($\|\nabla_\theta J(\hat{\theta})\|_2$ small in expectation — proof optional)

For convex function, this guarantees global optimality

Outline for today

✓ 1. Recap on Gradient descent and stochastic gradient descent

2. Warm up: computing gradient using importance weighting

3. Policy Gradient formulations

Warm Up: Importance Weighting

$$J(\theta) = \mathbb{E}_{x \sim P_\theta} \left[f(x) \right]$$

Δ

$$= \int_x P_\theta(x) f(x) \, dx$$

$P_\theta = N(\theta, \sigma^2 I)$ $\theta \in R^d$

$f : R^d \to R$

$$\nabla_\theta J(\theta) = \int_x \nabla_\theta P_\theta(x) f(x) \, dx$$

Warm Up: Importance Weighting

$$J(\theta) = \mathbb{E}_{x \sim P_\theta} \left[f(x) \right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$$

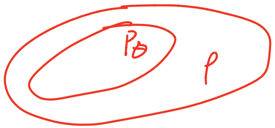
Warm Up: Importance Weighting

$$J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$$

$\rho \in \Delta(\mathcal{X})$

Suppose that I have a sampling distribution ρ, s.t., $\underbrace{\max_x P_\theta(x)/\rho(x)}_{c} < \infty$



Warm Up: Importance Weighting

$$J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x) = \nabla_\theta \mathbb{E}_{x \sim \rho} \frac{P_\theta(x)}{\rho(x)} f(x)$$

$$\mathbb{E}_{x \sim \rho} \frac{P_\theta(x)}{\rho(x)} = \int_x \rho(x) \frac{P_\theta(x)}{\rho(x)} = \int_x P_\theta(x) = \mathbb{E}_{x \sim P_\theta}$$

Warm Up: Importance Weighting

$$J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x) = \nabla_\theta \mathbb{E}_{x \sim \rho} \frac{P_\theta(x)}{\rho(x)} f(x) \quad = \mathbb{E}_{x \sim \rho} \frac{\nabla_\theta P_\theta(x)}{\rho(x)} f(x)$$

Warm Up: Importance Weighting

$$J(\theta) = \mathbb{E}_{x\sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x\sim P_\theta} f(x)$$

$\int_x \nabla_\theta P(x) f(x) dx$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x\sim P_\theta} f(x) = \nabla_\theta \mathbb{E}_{x\sim\rho} \frac{P_\theta(x)}{\rho(x)} f(x) = \mathbb{E}_{x\sim\rho} \frac{\nabla_\theta P_\theta(x)}{\rho(x)} f(x) \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\nabla_\theta P_\theta(x_i)}{\rho(x_i)} f(x_i)$$

unbiased-estimate

Let's sample $\{x_i\}_{i=1}^N \overset{iid}{\sim} \rho$

Warm Up: Importance Weighting

$$J(\theta) = \mathbb{E}_{x\sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x\sim P_\theta} f(x)$$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x\sim P_\theta} f(x) = \nabla_\theta \mathbb{E}_{x\sim \rho} \frac{P_\theta(x)}{\rho(x)} f(x) = \mathbb{E}_{x\sim\rho} \frac{\nabla_\theta P_\theta(x)}{\rho(x)} f(x) \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\nabla_\theta P_\theta(x_i)}{\rho(x_i)} f(x_i)$$

To compute gradient at θ_0: $\nabla_\theta J(\theta_0)$ (in short of $\nabla_\theta J(\theta)\,|_{\theta=\theta_0}$)

Warm Up: Importance Weighting

$$J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x) = \nabla_\theta \mathbb{E}_{x \sim \rho} \frac{P_\theta(x)}{\rho(x)} f(x) = \mathbb{E}_{x \sim \rho} \frac{\nabla_\theta P_\theta(x)}{\rho(x)} f(x) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\nabla_\theta P_\theta(x_i)}{\rho(x_i)} f(x_i)$$

To compute gradient at θ_0: $\nabla_\theta J(\theta_0)$ (in short of $\nabla_\theta J(\theta)\,|_{\theta=\theta_0}$)

We can set sampling distribution $\rho = P_{\theta_0}$ ← Available

Warm Up: Importance Weighting

$$J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$$

Suppose that I have a sampling distribution ρ, s.t., $\max_x P_\theta(x)/\rho(x) < \infty$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x) = \nabla_\theta \mathbb{E}_{x \sim \rho} \frac{P_\theta(x)}{\rho(x)} f(x) = \mathbb{E}_{x \sim \rho} \frac{\nabla_\theta P_\theta(x)}{\rho(x)} f(x) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\nabla_\theta P_\theta(x_i)}{\rho(x_i)} f(x_i)$$

To compute gradient at θ_0: $\nabla_\theta J(\theta_0)$ (in short of $\nabla_\theta J(\theta)|_{\theta=\theta_0}$)

We can set sampling distribution $\rho = P_{\theta_0}$

$$\nabla_\theta J(\theta_0) = \mathbb{E}_{x \sim P_{\theta_0}} \left[\nabla_\theta \ln P_{\theta_0}(x) \cdot f(x) \right]$$

by setting $\rho = P_{\theta_0}$: $\frac{\nabla_\theta P_{\theta_0}(x)}{P_{\theta_0}(x)}$

Sample $x \sim P_{\theta_0}$

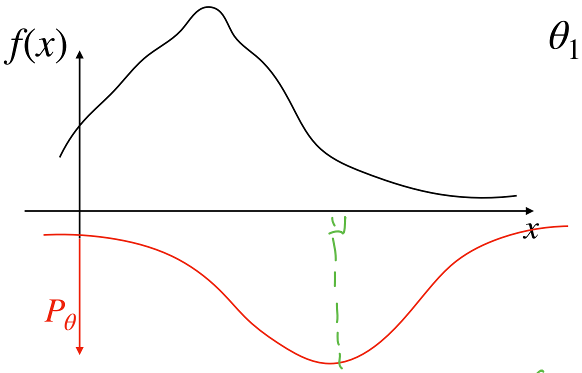
Compute: $\nabla_\theta \ln P_{\theta_0}(x) \cdot f(x)$

Warm Up

maximize

$$\nabla_\theta J(\theta)\,|_{\theta=\theta_0} = \mathbb{E}_{x\sim P_{\theta_0}} \nabla_\theta \ln P_{\theta_0}(x)\cdot f(x)$$

$$\theta_1 = \theta_0 + \eta \nabla_\theta J(\theta_0)$$

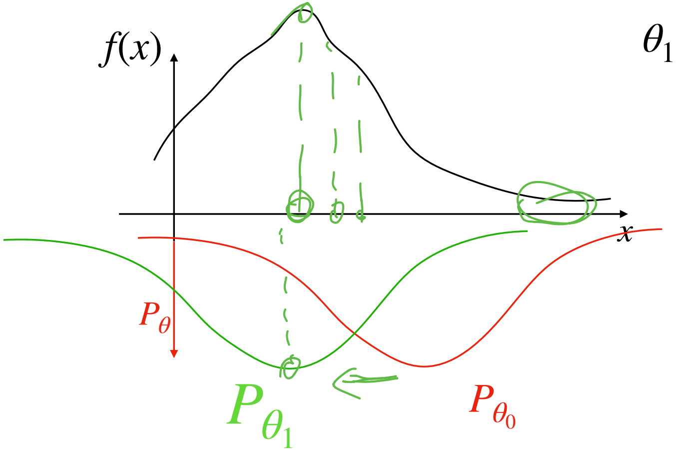


$P_{\theta_0} \rightarrow N(\theta_0, \sigma^2)$

Warm Up

$$\nabla_\theta J(\theta)\,|_{\theta=\theta_0} = \mathbb{E}_{x\sim P_{\theta_0}} \nabla_\theta \ln P_{\theta_0}(x) \cdot f(x)$$

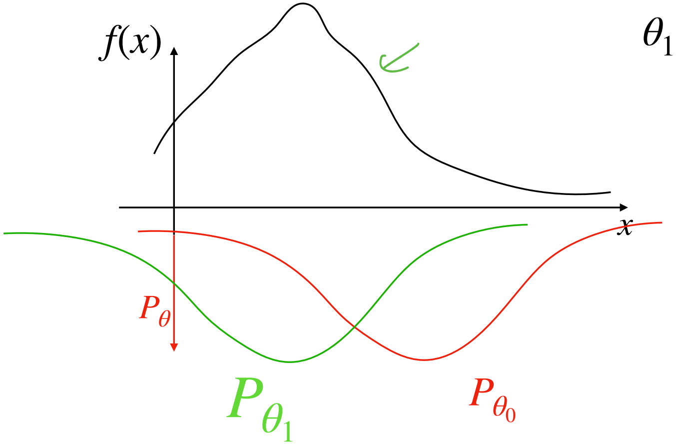
$$\theta_1 = \theta_0 + \eta \nabla_\theta J(\theta_0)$$



Warm Up

$$\nabla_\theta J(\theta)\,|_{\theta=\theta_0} = \mathbb{E}_{x\sim P_{\theta_0}} \nabla_\theta \ln P_{\theta_0}(x) \cdot f(x)$$

$$\theta_1 = \theta_0 + \eta \nabla_\theta J(\theta_0)$$



Update distribution (via updating θ) such that P_θ has high probability mass at regions where $f(x)$ is large

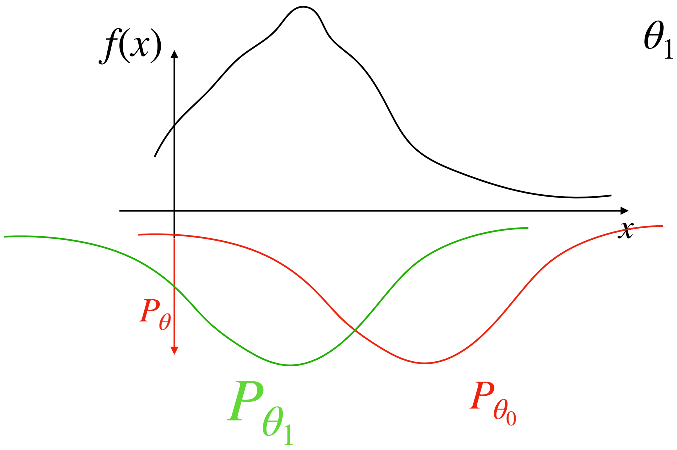
Warm Up

$$\nabla_\theta J(\theta)\,|_{\theta=\theta_0} = \mathbb{E}_{x\sim P_{\theta_0}} \nabla_\theta \ln P_{\theta_0}(x) \cdot f(x)$$

$$\theta_1 = \theta_0 + \eta \nabla_\theta J(\theta_0)$$

unbiased est:

$x_i \overset{i.i.d}{\sim} P_{\theta_0}$: $\frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \ln P_{\theta_0}(x_i) \cdot f(x_i)$



Update distribution (via updating θ) such that P_θ has high probability mass at regions where $f(x)$ is large

Using same idea, now let's move on to RL…

Outline for today

✓ 1. Recap on Gradient descent and stochastic gradient descent

✓ 2. Warm up: computing gradient using importance weighting

3. Policy Gradient formulations

Policy Gradient: Examples of Policy Parameterization (discrete actions)

Recall that we consider parameterized policy $\pi_\theta(\cdot \mid s) \in \Delta(A), \forall s$

1. Softmax Policy for discrete MDPs:

$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

Recall that we consider parameterized policy $\pi_\theta(\cdot \mid s) \in \Delta(A), \forall s$

1. Softmax Policy for discrete MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

$\pi_\theta(\cdot \mid s) \in \Delta(A)$

$\sum_a \pi_\theta(a \mid s) = 1$

$\pi_\theta(a \mid s) \geq 0, \forall a$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

Recall that we consider parameterized policy $\pi_\theta(\cdot \mid s) \in \Delta(A), \forall s$

1. Softmax Policy for discrete MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (We will try this in HW2)

Policy Gradient: Examples of Policy Parameterization (discrete actions)

Recall that we consider parameterized policy $\pi_\theta(\cdot \mid s) \in \Delta(A), \forall s$

1. Softmax Policy for discrete MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (We will try this in HW2)

Feature vector $\phi(s, a) \in \mathbb{R}^d$, and parameter $\theta \in \mathbb{R}^d$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

Recall that we consider parameterized policy $\pi_\theta(\cdot \mid s) \in \Delta(A), \forall s$

1. Softmax Policy for discrete MDPs:

$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (We will try this in HW2)

Feature vector $\phi(s, a) \in \mathbb{R}^d$, and parameter $\theta \in \mathbb{R}^d$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \phi(s, a))}{\sum_{a'} \exp(\theta^\top \phi(s, a'))}$$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

Recall that we consider parameterized policy $\pi_\theta(\cdot \mid s) \in \Delta(A), \forall s$

1. Softmax Policy for discrete MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (We will try this in HW2)

Feature vector $\phi(s, a) \in \mathbb{R}^d$, and parameter $\theta \in \mathbb{R}^d$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \phi(s, a))}{\sum_{a'} \exp(\theta^\top \phi(s, a'))}$$

3. Neural Policy:

Policy Gradient: Examples of Policy Parameterization (discrete actions)

Recall that we consider parameterized policy $\pi_\theta(\cdot \mid s) \in \Delta(A), \forall s$

1. Softmax Policy for discrete MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (We will try this in HW2)

Feature vector $\phi(s,a) \in \mathbb{R}^d$, and parameter $\theta \in \mathbb{R}^d$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \phi(s,a))}{\sum_{a'} \exp(\theta^\top \phi(s,a'))}$$

3. Neural Policy:

Neural network $f_\theta : S \times A \mapsto \mathbb{R}$

$\theta^*(s,a) \cdot c$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

Recall that we consider parameterized policy $\pi_\theta(\cdot \mid s) \in \Delta(A), \forall s$

1. Softmax Policy for discrete MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (We will try this in HW2)

Feature vector $\phi(s, a) \in \mathbb{R}^d$, and parameter $\theta \in \mathbb{R}^d$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \phi(s, a))}{\sum_{a'} \exp(\theta^\top \phi(s, a'))}$$

3. Neural Policy:

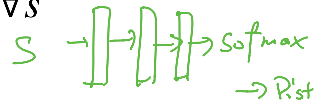
Neural network $f_\theta : S \times A \mapsto \mathbb{R}$

$$\pi_\theta(a \mid s) = \frac{\exp(f_\theta(s, a))}{\sum_{a'} \exp(f_\theta(s, a'))}$$

$f_{\theta^*} = c \cdot Q^*(s,a)$

Policy Gradient: Examples of Policy Parameterization (discrete actions)

Recall that we consider parameterized policy $\pi_\theta(\cdot \mid s) \in \Delta(A), \forall s$



1. Softmax Policy for discrete MDPs:

$$\theta_{s,a} \in \mathbb{R}, \forall s, a \in S \times A$$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

2. Softmax linear Policy (We will try this in HW2)

Feature vector $\phi(s, a) \in \mathbb{R}^d$, and parameter $\theta \in \mathbb{R}^d$

$$\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \phi(s, a))}{\sum_{a'} \exp(\theta^\top \phi(s, a'))}$$

3. Neural Policy:

Neural network $f_\theta : S \times A \mapsto \mathbb{R}$

$$\pi_\theta(a \mid s) = \frac{\exp(f_\theta(s, a))}{\sum_{a'} \exp(f_\theta(s, a'))}$$

$\nabla \ln \pi_\theta(a \mid s)$

In high level, think about π_θ as a classifier which has its parameters to be optimized

Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \ldots\}$$

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

$\pi_\theta \rightarrow \rho_\theta$

Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \ldots\}$$

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

← Markov property

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)}\underbrace{\left[\sum_{h=0}^{\infty}\gamma^h r(s_h, a_h)\right]}_{R(\tau)}$$

$$R(\tau) = \sum_{h=0}^{\infty} \gamma^h r(s_h, a_h)$$

$$\mathbb{E}_{x\sim P_\theta}[f(x)]$$

$$J(\theta) = \mathbb{E}_{\tau\sim P_\theta(\tau)}[R(\tau)]$$

Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \ldots\}$$

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

$$J(\pi_\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \right]}_{R(\tau)}$$

$$\nabla_\theta J(\pi_{\theta_0}) = \mathbb{E}_{\tau \sim \rho_{\theta_0}(\tau)} \left[\nabla_\theta \ln \rho_{\theta_0}(\tau) R(\tau) \right]$$

← Importance Weighting Trick.

Recall $P(s' \mid s, a)$ ← unknown

$$\nabla_\theta \left[\mathbb{E}_{x \sim P_\theta} [f(x)] \right] \Big|_{\theta = \theta_0} = \mathbb{E}_{x \sim P_{\theta_0}} \nabla_\theta \ln P_{\theta_0}(x) \cdot f(x)$$

Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \ldots\}$$

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

$$\rightarrow \ln(\rho_\theta(\tau)) = \ln \mu(s_0) + \ln \pi_\theta(a_0 \mid s_1) + \ln P(s_1 \mid s_0, a_0) + \ldots$$

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)}\underbrace{\left[\sum_{h=0}^{\infty}\gamma^h r(s_h, a_h)\right]}_{R(\tau)}$$

$$\nabla_\theta \ln(\rho_\theta(\tau)) = \nabla_\theta \ln \mu(s_0) + \nabla_\theta \ln \pi_\theta(a_0 \mid s_0) + \nabla_\theta \ln P(s_1 \mid s_0, a_0) + \ldots$$

$\nabla_\theta \ln \mu(s_0) = 0$, $\nabla_\theta \ln P(s_1 \mid s_0, a_0) = 0$

$$\nabla_\theta J(\pi_{\theta_0}) = \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)}\left[\nabla_\theta \ln \rho_{\theta_0}(\tau) R(\tau)\right]$$

$$= \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)}\left[\nabla_\theta\left(\ln \rho(s_0) + \ln \pi_{\theta_0}(a_0 \mid s_0) + \ln P(s_1 \mid s_0, a_0) + \ldots\right) R(\tau)\right]$$

($\ln \rho(s_0)$: $=0$; $\ln P(s_1 \mid s_0, a_0)$: $=0$)

Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \ldots\}$$

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h)\right]}_{R(\tau)}$$

$$\nabla_\theta J(\pi_{\theta_0}) = \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)} \left[\nabla_\theta \ln \rho_{\theta_0}(\tau) R(\tau)\right]$$

$$= \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)} \left[\nabla_\theta \left(\ln \rho(s_0) + \ln \pi_{\theta_0}(a_0 \mid s_0) + \ln P(s_1 \mid s_0, a_0) + \ldots\right) R(\tau)\right]$$

$$= \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)} \left[\nabla_\theta \left(\ln \pi_{\theta_0}(a_0 \mid s_0) + \ln \pi_{\theta_0}(a_1 \mid s_1)\ldots\right) R(\tau)\right]$$

Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \ldots\}$$

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h)\right]}_{R(\tau)}$$

$$\nabla_\theta J(\pi_{\theta_0}) = \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)} \left[\nabla_\theta \ln \rho_{\theta_0}(\tau) R(\tau)\right]$$

$\nabla_\theta \ln \pi_\theta(a|s)$

$$= \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)} \left[\nabla_\theta \left(\ln \mu(s_0) + \ln \pi_{\theta_0}(a_0 \mid s_0) + \ln P(s_1 \mid s_0, a_0) + \ldots\right) R(\tau)\right]$$

$=0 \quad =0$

$$= \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)} \left[\nabla_\theta \left(\ln \pi_{\theta_0}(a_0 \mid s_0) + \ln \pi_{\theta_0}(a_1 \mid s_1)\ldots\right) R(\tau)\right] = \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)} \left[\left(\sum_{h=0}^{\infty} \nabla_\theta \ln \pi_{\theta_0}(a_h \mid s_h)\right) R(\tau)\right]$$

Derivation of Policy Gradient: REINFORCE

$$\tau = \{s_0, a_0, s_1, a_1, \ldots\}$$

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots$$

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)}\underbrace{\left[\sum_{h=0}^{\infty}\gamma^h r(s_h, a_h)\right]}_{R(\tau)}$$

Adjust policy such that larger reward traj has higher likelihood

$$\nabla_\theta J(\pi_{\theta_0}) = \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)}\left[\nabla_\theta \ln \rho_{\theta_0}(\tau)R(\tau)\right]$$

$$= \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)}\left[\nabla_\theta\left(\ln \rho(s_0) + \ln \pi_{\theta_0}(a_0 \mid s_0) + \ln P(s_1 \mid s_0, a_0) + \ldots\right)R(\tau)\right]$$

$$= \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)}\left[\nabla_\theta\left(\ln \pi_{\theta_0}(a_0 \mid s_0) + \ln \pi_{\theta_0}(a_1 \mid s_1)\ldots\right)R(\tau)\right] = \mathbb{E}_{\tau\sim\rho_{\theta_0}(\tau)}\left[\left(\sum_{h=0}^{\infty}\nabla_\theta \ln \pi_{\theta_0}(a_h \mid s_h)\right)R(\tau)\right]$$

Summary so far for Policy Gradients

We derived the most classic PG formulation:

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \left[\left(\sum_{h=0}^{\infty} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \right) R(\tau) \right]$$

$$\tau \sim \rho_\theta(\tau)$$

$$\sum_{h=0}^{\infty} \nabla_\theta \ln \pi_\theta (a_h \mid s_h) \boxed{R(\tau)}$$

$$= \sum_{h=0}^{\infty} \gamma^h r(s_h, a_h)$$

Summary so far for Policy Gradients

We derived the most classic PG formulation:

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \left[\left(\sum_{h=0}^{\infty} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \right) R(\tau) \right]$$

Increase the likelihood of sampling an trajectory with high total reward

For finite horizon MDP (very common setting for PG):

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \left[\left(\sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \right) R(\tau) \right]$$

where $R(\tau) = \sum_{h=0}^{H-1} r(s_h, a_h)$

$\tau \sim \rho_\theta(\tau)$ $\quad \sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \left[\sum_{t=0}^{H-1} r(s_t, a_t) \right]$

For finite horizon MDP (very common setting for PG):

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \left[\left(\sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \right) R(\tau) \right]$$

where $R(\tau) = \sum_{h=0}^{H-1} r(s_h, a_h)$

Increase the likelihood of sampling an trajectory with high total reward

Further simplification on PG (e.g., for finite horizon)

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \left[\left(\sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \cdot \sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \right) \right]$$

Further simplification on PG (e.g., for finite horizon)

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \left[\left(\sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h | s_h) \cdot \sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \right) \right]$$

(Change action distribution at h only affects rewards later on…)

Further simplification on PG (e.g., for finite horizon)

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \left[\left(\sum_{h=0}^{H-1} \nabla_\theta \ln \pi_\theta(a_h \mid s_h) \cdot \sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \right) \right]$$

(Change action distribution at h only affects rewards later on…)

Exercise:

Show this simplified version is equivalent to REINFORCE

Summary for today

1. Importance Weighting Trick ✓

2. Policy Gradient:

REINFORCE (a direct application of our warm up example):

Summary for today

1. Importance Weighting Trick

2. Policy Gradient:

REINFORCE (a direct application of our warm up example):

$$\nabla J(\theta_t) = \mathbb{E}_{\tau \sim \rho_{\theta_t}(\tau)} \left[\left(\sum_{h=0}^{H-1} \nabla_\theta \ln \pi_{\theta_t}(a_h \mid s_h) \right) R(\tau) \right]$$

Summary for today

1. Importance Weighting Trick

2. Policy Gradient:

REINFORCE (a direct application of our warm up example):

$$\nabla J(\theta_t) = \mathbb{E}_{\tau \sim \rho_{\theta_t}(\tau)} \left[\left(\sum_{h=0}^{H-1} \nabla_\theta \ln \pi_{\theta_t}(a_h \mid s_h) \right) R(\tau) \right]$$

$E[g_t] = \nabla J(\theta_t)$ $\theta_{t+1} = \theta_t + \eta \cdot g_t$

2. Use unbiased estimate of $\nabla_\theta J(\theta)$, SG ascent converges to local optimal policy

$$\hat{A}^t(s,a) \approx A^{\pi^t}(s,a), \ \forall a$$

$$\arg\max_a \hat{A}^t(s,a) \neq \pi^t(s)$$

$s \sim d_\mu^{\pi^t}$, $a \sim U(A)$. y s.t. $E[y] = A^{\pi^t}(s,a)$

$$\hat{A} = \arg\min_f \sum \left(f(s,a) - y \right)^2$$

$$\pi^{t+1}(s) = \arg\max_a \hat{A}(s,a)$$

$\hat{A}$ function Approximate

$$\arg\max_a A^{\pi}(s,a) = \arg\max_a Q^{\pi}(s,a)$$

$= Q^{\pi}(s,a) - V^{\pi}(s)$