Policy Gradient Methods (continued)

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning

Fall 2022

Today

- Recap++
  - will clarify few points based on feedback.

- Today:
  1. Estimation of Stochastic Gradients
  2. Variance Reduction
  3. More Variance Reduction (baselines)

Recap++

(some new material and clarifications)

Recap Outline:

1. The Learning Setting
2. Objective: direct policy optimization.
3. General convergence: properties of SGD
4. Importance Sampling
   & Deriving a Policy Gradient Expression

The Learning Setting:

We don’t know the MDP, but we can obtain trajectories.

The Learning Setting:

We don't know the MDP, but we can obtain trajectories.

The Finite Horizon, Learning Setting. We can obtain trajectories as follows:

- We start at $s_0 \sim \mu_0$.
- We act for H steps and observe the trajectory $\tau = \{s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}\}$

The Learning Setting:

We don't know the MDP, but we can obtain trajectories.

The Finite Horizon, Learning Setting. We can obtain trajectories as follows:

- We start at $s_0 \sim \mu_0$.
- We act for H steps and observe the trajectory $\tau = \{s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}\}$

The Infinite Horizon, Discounted Learning Setting. We can obtain trajectories as follows:

- We start at $s_0 \sim \mu_0$.
- We can obtain a "long trajectories" $\tau = \{s_0, a_0, s_1, a_1, \ldots\}$
  - Suppose we can terminate the trajectory at will.
    (and sufficient long trajectories will well approximate the discounted value function)

The Learning Setting:

We don't know the MDP, but we can obtain trajectories.

The Finite Horizon, Learning Setting. We can obtain trajectories as follows:

- We start at $s_0 \sim \mu_0$.
- We act for H steps and observe the trajectory $\tau = \{s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}\}$

The Infinite Horizon, Discounted Learning Setting. We can obtain trajectories as follows:

- We start at $s_0 \sim \mu_0$.
- We can obtain a "long trajectories" $\tau = \{s_0, a_0, s_1, a_1, \ldots\}$
  - Suppose we can terminate the trajectory at will.
    (and sufficient long trajectories will well approximate the discounted value function)

Note that with a simulator, we can sample trajectories as specified in the above.

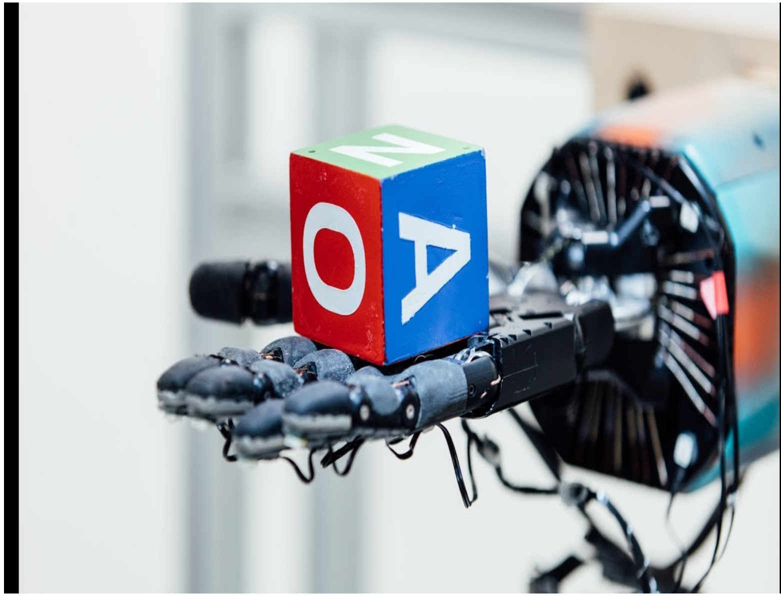
Recap Outline:

1. The Learning Setting
2. Objective: direct policy optimization.
3. General convergence: properties of SGD
4. Importance Sampling
   & Deriving a Policy Gradient Expression

Policy Optimization:
our goal is to do well on “large” problems

[AlphaZero, Silver et.al, 17]

[OpenAI Five, 18]



[OpenAI,19]

Recap: Policy Parameterization

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

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

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

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

2. Neural Policy:

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

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

Our Objective and Policy Gradient Ascent

We consider either discounted or finite horizon settings.

$$J(\theta) := E_{s_0 \sim \mu_0}\left[V^{\pi_\theta}(s_0)\right] = E\left[\sum_{h=0}^{\infty} \gamma^h r_h \middle| \mu_0, \pi_\theta\right]$$

$$J(\theta) := E_{s_0 \sim \mu_0}\left[V^{\pi_\theta}(s_0)\right] = E\left[\sum_{h=0}^{H-1} r_h \middle| \mu_0, \pi_\theta\right]$$

- Objective: try to find "good" parameters

$$\max_{\theta} J(\theta)$$

ascent,

- Approach: stochastic gradient ~~descent~~ (or gradient ~~descent~~)

$$\theta_{t+1} = \theta_t + \eta \widetilde{\nabla}_\theta J(\theta_t)$$

$$E[\widehat{\nabla}_\theta J(\theta_t)] = \nabla J(\theta_t)$$

Recap Outline:

1. The Learning Setting
2. Objective: direct policy optimization.
3. General convergence: properties of SGD
4. Importance Sampling
   & Deriving a Policy Gradient Expression

Stochastic Gradient Descent

$(x,y) \sim D$

$\tilde{\nabla} J(\theta_t) = 2(f_\theta(x)-y)\nabla f_\theta(x)$

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

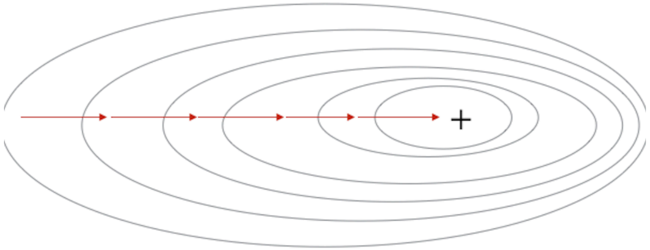
SGD minimizes the above objective function as follows:

Initialize θ_0, for t = 0, … :

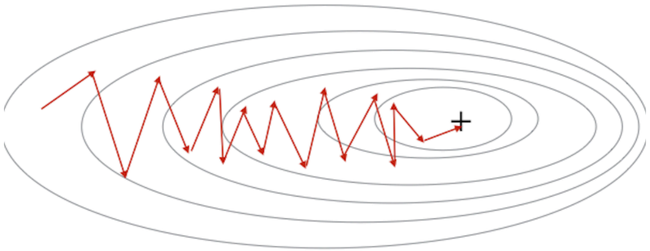
$$\theta_{t+1} = \theta_t - \eta_t \widetilde{\nabla}_\theta J(\theta_t)$$

where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t)$

Gradient Descent



Stochastic Gradient Descent



SGD: Convergence to a Stationary Point for Nonconvex Functions

- Def of β-smooth: $\|\nabla_\theta J(\theta) - \nabla_\theta J(\theta_0)\|_2 \leq \beta \|\theta - \theta_0\|_2$

SGD: Convergence to a Stationary Point for Nonconvex Functions

- Def of β-smooth: $\|\nabla_\theta J(\theta) - \nabla_\theta J(\theta_0)\|_2 \le \beta \|\theta - \theta_0\|_2$

- [**Theorem**] (informal) Suppose we run SGD: $\theta_{t+1} = \theta_t - \eta \widetilde{\nabla}_\theta J(\theta_t)$, for T steps, where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t)$ with $\eta = O(1/\sqrt{T})$. Assume:
  - $J(\theta)$ is β-smooth.
  - $J(\theta)$ is bounded: $|J(\theta)| \le M, \quad \forall \theta$.
  - $\widetilde{\nabla}_\theta J(\theta)$ has "bounded variance": $\mathbb{E}\left[\|\nabla_\theta J(\theta_t) - \widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \le \sigma^2$,

then, in T steps, SGD will find a θ such that:

$$\|\nabla_\theta J(\theta)\| \le O\left(\left(M\beta\sigma^2/T\right)^{1/4} + \left(M\beta/T\right)^{1/2}\right).$$

SGD: Convergence to a Stationary Point for Nonconvex Functions

- Def of β-smooth: $\|\nabla_\theta J(\theta) - \nabla_\theta J(\theta_0)\|_2 \le \beta \|\theta - \theta_0\|_2$

- [**Theorem**] (informal) Suppose we run SGD: $\theta_{t+1} = \theta_t - \eta \widetilde{\nabla}_\theta J(\theta_t)$, for T steps, where $\mathbb{E}\left[\widetilde{\nabla}_\theta J(\theta_t)\right] = \nabla_\theta J(\theta_t)$ with $\eta = O(1/\sqrt{T})$. Assume:
  - $J(\theta)$ is β-smooth.
  - $J(\theta)$ is bounded: $|J(\theta)| \le M, \quad \forall \theta.$
  - $\widetilde{\nabla}_\theta J(\theta)$ has "bounded variance": $\mathbb{E}\left[\|\nabla_\theta J(\theta_t) - \widetilde{\nabla}_\theta J(\theta_t)\|_2^2\right] \le \sigma^2$,

then, in T steps, SGD will find a θ such that:

$$\|\nabla_\theta J(\theta)\| \le O\left(\left(M\beta\sigma^2/T\right)^{1/4} + \left(M\beta/T\right)^{1/2}\right).$$

Formally, we have $\mathbb{E}\left[\min_{t \le T} \|\nabla_\theta J(\theta_t)\|^2\right] \le O\left(\sqrt{M\beta\sigma^2/T} + M\beta/T\right)$

Recap Outline:

1. The Learning Setting
2. Objective: direct policy optimization.
3. General convergence: properties of SGD
4. Importance Sampling
   & Deriving a Policy Gradient Expression

$$\mathbb{E}_{x \sim p}[f(x)] = \mathbb{E}_{x \sim p}\left[\frac{P(x)}{p(x)} f(x)\right]$$

$$= \sum_x P(x) f(x) = \sum_x p(x) \frac{P(x)}{p(x)} f(x)$$

Why is this helpful?

- Variance reduction

- our data collection was under p $\quad x_i \sim p$

(say we know P) $\quad f(x_i)$

Importance Sampling (and the Likelihood Ratio Method)

Importance Sampling (and the Likelihood Ratio Method)

For $J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$, our goal is to accurately approximate $\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$.
(We want to avoid computing the integral/sum.)

Importance Sampling (and the Likelihood Ratio Method)

For $J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$, our goal is to accurately approximate $\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$.
(We want to avoid computing the integral/sum.)

- Often, we are in setting where:
  - P_θ is "easy" to compute.
  - We have a distribution ρ, that is easy to sample from and where $\max_x P_\theta(x)/\rho(x) < \infty$

    (sometimes we use P_θ itself as ρ)

Importance Sampling (and the Likelihood Ratio Method)

For $J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$, our goal is to accurately approximate $\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$.
(We want to avoid computing the integral/sum.)

- Often, we are in setting where:
  - P_θ is "easy" to compute.
  - We have a distribution ρ, that is easy to sample from and where $\max_x P_\theta(x)/\rho(x) < \infty$

    (sometimes we use P_θ itself as ρ)

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

Importance Sampling (and the Likelihood Ratio Method)

For $J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$, our goal is to accurately approximate $\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$.
(We want to avoid computing the integral/sum.)

- Often, we are in setting where:
  - P_θ is "easy" to compute.
  - We have a distribution ρ, that is easy to sample from and where $\max_x P_\theta(x)/\rho(x) < \infty$

    (sometimes we use P_θ itself as ρ)

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

Importance Sampling (and the Likelihood Ratio Method)

For $J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$, our goal is to accurately approximate $\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$.
(We want to avoid computing the integral/sum.)

- Often, we are in setting where:
  - P_θ is "easy" to compute.
  - We have a distribution ρ, that is easy to sample from and where $\max_x P_\theta(x)/\rho(x) < \infty$

    (sometimes we use P_θ itself as ρ)

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

Importance Sampling (and the Likelihood Ratio Method)

For $J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$, our goal is to accurately approximate $\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$.
(We want to avoid computing the integral/sum.)

- Often, we are in setting where:
  - P_θ is "easy" to compute.
  - We have a distribution ρ, that is easy to sample from and where $\max_x P_\theta(x)/\rho(x) < \infty$

    (sometimes we use P_θ itself as ρ)

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

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

Importance Sampling (and the Likelihood Ratio Method)

For $J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$, our goal is to accurately approximate $\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$.
(We want to avoid computing the integral/sum.)

- Often, we are in setting where:
  - P_θ is "easy" to compute.
  - We have a distribution ρ, that is easy to sample from and where $\max_x P_\theta(x)/\rho(x) < \infty$

    (sometimes we use P_θ itself as ρ)

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

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

By setting the sampling distribution $\rho = P_{\theta_0}$

Importance Sampling (and the Likelihood Ratio Method)

For $J(\theta) = \mathbb{E}_{x \sim P_\theta}\left[f(x)\right]$, our goal is to accurately approximate $\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{x \sim P_\theta} f(x)$.
(We want to avoid computing the integral/sum.)

- Often, we are in setting where:
  - P_θ is "easy" to compute.
  - We have a distribution ρ, that is easy to sample from and where $\max_x P_\theta(x)/\rho(x) < \infty$

    (sometimes we use P_θ itself as ρ)

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

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

By setting the sampling distribution $\rho = P_{\theta_0}$

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

Recap: the REINFORCE Algorithm (discounted case)

We derived the most basic PG formulation:

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

Recap: the REINFORCE Algorithm (discounted case)

We derived the most basic PG formulation:

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

Increase the likelihood of sampling an trajectory with high total reward

Recap: the REINFORCE Algorithm

(finite horizon case)

$$\tau = \{s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}\}$$

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

Recap: the REINFORCE Algorithm

(finite horizon case)

$$\tau = \{s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}\}$$

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

$$J(\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{H-1} r(s_h, a_h)\right]}_{R(\tau)}$$

Recap: the REINFORCE Algorithm

(finite horizon case)

$$\tau = \{s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}\}$$

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

$$J(\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{H-1} r(s_h, a_h)\right]}_{R(\tau)}$$

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

Derivation of Policy Gradient: REINFORCE for finite H

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

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots P(s_{H-1} \mid s_{H-2}, a_{H-2})\pi_\theta(a_{H-1} \mid s_{H-1})$$

Derivation of Policy Gradient: REINFORCE for finite H

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

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots P(s_{H-1} \mid s_{H-2}, a_{H-2})\pi_\theta(a_{H-1} \mid s_{H-1})$$

$$J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{H-1} r(s_h, a_h) \right]}_{R(\tau)}$$

Derivation of Policy Gradient: REINFORCE for finite H

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

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\dots P(s_{H-1} \mid s_{H-2}, a_{H-2})\pi_\theta(a_{H-1} \mid s_{H-1})$$

$$J(\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{H-1} r(s_h, a_h)\right]}_{R(\tau)}$$

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

Derivation of Policy Gradient: REINFORCE for finite H

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

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots P(s_{H-1} \mid s_{H-2}, a_{H-2})\pi_\theta(a_{H-1} \mid s_{H-1})$$

$$J(\theta) = \mathbb{E}_{\tau\sim\rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{H-1} r(s_h, a_h)\right]}_{R(\tau)}$$

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

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

Derivation of Policy Gradient: REINFORCE for finite H

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

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots P(s_{H-1} \mid s_{H-2}, a_{H-2})\pi_\theta(a_{H-1} \mid s_{H-1})$$

$$J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{H-1} r(s_h, a_h) \right]}_{R(\tau)}$$

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

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

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

Derivation of Policy Gradient: REINFORCE for finite H

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

$$\rho_\theta(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)\ldots P(s_{H-1} \mid s_{H-2}, a_{H-2})\pi_\theta(a_{H-1} \mid s_{H-1})$$

$$J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \underbrace{\left[\sum_{h=0}^{H-1} r(s_h, a_h) \right]}_{R(\tau)}$$

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

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

$$= \mathbb{E}_{\tau \sim \rho_{\theta_0}(\tau)} \left[\nabla_\theta \left(\ln \pi_{\theta_0}(a_0 \mid s_0) + \ln \pi_{\theta_0}(a_1 \mid s_1) \ldots \right) R(\tau) \right] = \mathbb{E}_{\tau \sim \rho_{\theta_0}(\tau)} \left[\left(\sum_{h=0}^{H-1} \nabla_\theta \ln \pi_{\theta_0}(a_h \mid s_h) \right) R(\tau) \right]$$

Today:

Policy Gradient Descent

Outline:

1. Estimation of Stochastic Gradients
2. Variance Reduction
3. More Variance Reduction (baselines)

Obtaining an Unbiased Gradient Estimate at θ_0

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

Obtaining an Unbiased Gradient Estimate at θ_0

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

1. Obtain a trajectory $\tau \sim \rho_{\theta_0}$
   (which we can do in our learning setting)

Obtaining an Unbiased Gradient Estimate at θ_0

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

$$R(\tau) = \sum_{h=0}^{H-1} r_h$$

1. Obtain a trajectory $\tau \sim \rho_{\theta_0}$
   (which we can do in our learning setting)
2. Set:

$$\widetilde{\nabla}_\theta J(\theta_0) := \sum_{h=0}^{H-1}\nabla \ln \pi_{\theta_0}(a_h|s_h)R(\tau)$$

Obtaining an Unbiased Gradient Estimate at θ_0

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

1. Obtain a trajectory $\tau \sim \rho_{\theta_0}$

   (which we can do in our learning setting)
2. Set:

$$\widetilde{\nabla}_\theta J(\theta_0) := \sum_{h=0}^{H-1} \nabla \ln \pi_{\theta_0}(a_h | s_h) R(\tau)$$

We have: $\mathbb{E}[\widetilde{\nabla}_\theta J(\theta_0)] = \nabla_\theta J(\pi_{\theta_0})$

PG with REINFORCE:

PG with REINFORCE:

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$

PG with REINFORCE:

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :

PG with REINFORCE:

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :
    1. Obtain a trajectory $\tau \sim \rho_{\theta_t}$

       Set $\widetilde{\nabla}_\theta J(\theta_t) = \sum_{h=0}^{H-1} \nabla \ln \pi_{\theta_t}(a_h | s_h) R(\tau)$

PG with REINFORCE:

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :
   1. Obtain a trajectory $\tau \sim \rho_{\theta_t}$

      Set $\widetilde{\nabla}_\theta J(\theta_t) = \sum_{h=0}^{H-1} \nabla \ln \pi_{\theta_t}(a_h | s_h) R(\tau)$

   2. Update: $\theta_{t+1} = \theta_t + \eta_t \widetilde{\nabla}_\theta J(\theta_t)$

The (mini-batch) PG procedure with REINFORCE

(reducing variance using batch sizes of M)

The (mini-batch) PG procedure with REINFORCE

(reducing variance using batch sizes of M)

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$

The (mini-batch) PG procedure with REINFORCE

(reducing variance using batch sizes of M)

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :

The (mini-batch) PG procedure with REINFORCE

(reducing variance using batch sizes of M)

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :
    1. Init $g = 0$ and do M times:

       Obtain a trajectory $\tau \sim \rho_{\theta_t}$

       Set $g = g + \sum_{h=0}^{H-1} \nabla \ln \pi_{\theta_t}(a_h | s_h) R(\tau)$

       Set $\widetilde{\nabla}_\theta J(\theta_t) := \frac{1}{M} g$

The (mini-batch) PG procedure with REINFORCE

(reducing variance using batch sizes of M)

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :
    1. Init $g = 0$ and do M times:

       Obtain a trajectory $\tau \sim \rho_{\theta_t}$

       Set $g = g + \sum_{h=0}^{H-1} \nabla \ln \pi_{\theta_t}(a_h | s_h) R(\tau)$

       Set $\widetilde{\nabla}_\theta J(\theta_t) := \frac{1}{M} g$

    2. Update: $\theta_{t+1} = \theta_t + \eta_t \widetilde{\nabla}_\theta J(\theta_t)$

Outline:

1. Estimation of Stochastic Gradients
2. Variance Reduction
3. More Variance Reduction: Baselines and Advantages

$\nabla_\theta \log \pi_\theta(a_{H-1} | s_{H-1}) R(\tau)$

A improved PG formulation, for sampling (finite horizon setting)

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

A improved PG formulation, for sampling (finite horizon setting)

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

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

A improved PG formulation, for sampling (finite horizon setting)

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

REINFORCE

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

nice for sampling

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

"usual"

A improved PG formulation, for sampling (finite horizon setting)

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

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

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

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

A improved PG formulation, for sampling (finite horizon setting)

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

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

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

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

HW: You will show these simplified version are also valid PG expressions

Proof sketch

$$\mathbb{E}_{x\sim P_\theta}\left[\nabla \log P_\theta(x)\right] = \sum_x P_\theta(x)\,\nabla \log P_\theta(x) = \sum_x \cancel{P_\theta(x)}\,\frac{\nabla P_\theta(x)}{\cancel{P_\theta(x)}}$$

$$= \sum_x \nabla P_\theta(x) = \nabla \sum_x P_\theta(x) = \nabla 1 = 0$$

Proof sketch

Let $f(s_0, a_0, \ldots s_{h-1}, a_{h-1}, s_h)$ be an arbitrary function.

$$\mathbb{E}_{a_h \sim \pi_\theta(\cdot|s_h)} \left[\nabla_\theta \ln \pi_\theta(a_h \mid s_h) f(s_0, a_0, \ldots s_{h-1}, a_{h-1}, s_h) \middle| s_0, a_0, \ldots s_{h-1}, a_{h-1}, s_h \right]$$

Proof sketch

Let $f(s_0, a_0, \ldots s_{h-1}, a_{h-1}, s_h)$ be an arbitrary function.

$$\mathbb{E}_{a_h \sim \pi_\theta(\cdot|s_h)} \left[\nabla_\theta \ln \pi_\theta(a_h \mid s_h) f(s_0, a_0, \ldots s_{h-1}, a_{h-1}, s_h) \middle| s_0, a_0, \ldots s_{h-1}, a_{h-1}, s_h \right]$$

$$= f(s_0, a_0, \ldots s_{h-1}, a_{h-1}, s_h) \mathbb{E}_{a_h \sim \pi_\theta(\cdot|s_h)} \left[\nabla_\theta \ln \pi_\theta(a_h \mid s_h) \middle| s_0, a_0, \ldots s_{h-1}, a_{h-1}, s_h \right] = \cancel{??} \quad 0$$

An improved PG procedure:

An improved PG procedure:

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$

An improved PG procedure:

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :

An improved PG procedure:

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :
    1. Obtain a trajectory $\tau \sim \rho_{\theta_t}$

       Set $\widetilde{\nabla}_\theta J(\theta_t) = \sum_{h=0}^{H-1} \left(\nabla \ln \pi_{\theta_t}(a_h \mid s_h) \sum_{t=h}^{H-1} r_t \right)$

An ~~improved~~ "usual" PG procedure:

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :
   1. Obtain a trajectory $\tau \sim \rho_{\theta_t}$

      Set $\widetilde{\nabla}_\theta J(\theta_t) = \sum_{h=0}^{H-1} \left(\nabla \ln \pi_{\theta_t}(a_h \mid s_h) \sum_{t=h}^{H-1} r_t \right)$

   2. Update: $\theta_{t+1} = \theta_t + \eta_t \widetilde{\nabla}_\theta J(\theta_t)$

A improved PG formulation, for sampling (for the discounted setting)

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

A improved PG formulation, for sampling (for the discounted setting)

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

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

A improved PG formulation, for sampling (for the discounted setting)

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

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

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

Outline:

1. Estimation of Stochastic Gradients
2. Variance Reduction
3. More Variance Reduction: Baselines and Advantages

With a “baseline” function:

With a "baseline" function:

For any function $b_h(s)$, we have:

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

With a “baseline” function:

did n't have this earlier.

For any function $b_h(s)$, we have:

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

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

(M=1) SGD with a Naive (constant) Baseline:

(M=1) SGD with a Naive (constant) Baseline:

- On a trajectory τ, define:

$$R_h(\tau) = \sum_{t=h}^{H-1} r_t.$$

(M=1) SGD with a Naive (constant) Baseline:

- On a trajectory τ, define:
  $$R_h(\tau) = \sum_{t=h}^{H-1} r_t.$$
- Let try to use a constant (time-dependent) baseline:
  $$b_h^{\theta} = \mathbb{E}_{\tau \sim \rho_{\theta}(\tau)} E\left[R_h(\tau)\right]$$

(M=1) SGD with a Naive (constant) Baseline:

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$

- On a trajectory τ, define:
  $$R_h(\tau) = \sum_{t=h}^{H-1} r_t.$$
- Let try to use a constant (time-dependent) baseline:
  $b_h^{\theta} = \mathbb{E}_{\tau \sim \rho_{\theta}(\tau)} E\left[R_h(\tau)\right]$

(M=1) SGD with a Naive (constant) Baseline:

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :

- On a trajectory τ, define:
  $$R_h(\tau) = \sum_{t=h}^{H-1} r_t.$$
- Let try to use a constant (time-dependent) baseline:
  $$b_h^{\theta} = \mathbb{E}_{\tau \sim \rho_{\theta}(\tau)} E\left[R_h(\tau)\right]$$

(M=1) SGD with a Naive (constant) Baseline:

- On a trajectory τ, define:

$$R_h(\tau) = \sum_{t=h}^{H-1} r_t.$$

- Let try to use a constant (time-dependent) baseline:

$$b_h^{\theta} = \mathbb{E}_{\tau \sim \rho_{\theta}(\tau)} E\left[R_h(\tau)\right]$$

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :
    1. Using N trajectories sampled under π_{θ_t}, set

$$\widetilde{b}_h = \frac{1}{N}\sum_{i=1}^{N} R_h(\tau_i)$$

(M=1) SGD with a Naive (constant) Baseline:

- On a trajectory τ, define:

$$R_h(\tau) = \sum_{t=h}^{H-1} r_t.$$

- Let try to use a constant (time-dependent) baseline: $b_h^{\theta} = \mathbb{E}_{\tau \sim \rho_{\theta}(\tau)} E\left[R_h(\tau)\right]$

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :
   1. Using N trajectories sampled under π_{θ_t}, set

      $$\widetilde{b}_h = \frac{1}{N} \sum_{i=1}^{N} R_h(\tau_i)$$

   2. Obtain a trajectory $\tau \sim \rho_{\theta_t}$

      Set $\widetilde{\nabla}_{\theta} J(\theta_t) = \sum_{h=0}^{H-1} \nabla \ln \pi_{\theta_t}(a_h \mid s_h) \left(R_h(\tau) - \widetilde{b}_h \right)$

(M=1) SGD with a Naive (constant) Baseline:

- On a trajectory τ, define:

$$R_h(\tau) = \sum_{t=h}^{H-1} r_t.$$

- Let try to use a constant (time-dependent) baseline: $b_h^{\theta} = \mathbb{E}_{\tau\sim\rho_\theta(\tau)} E\left[R_h(\tau)\right]$

1. Initialize θ_0, parameters: $\eta_1, \eta_2, \ldots$
2. For t = 0, … :
   1. Using N trajectories sampled under π_{θ_t}, set
   $$\widetilde{b}_h = \frac{1}{N}\sum_{i=1}^{N} R_h(\tau_i)$$
   2. Obtain a trajectory $\tau \sim \rho_{\theta_t}$
   Set $\widetilde{\nabla}_\theta J(\theta_t) = \sum_{h=0}^{H-1} \nabla \ln \pi_{\theta_t}(a_h | s_h)\left(R_h(\tau) - \widetilde{b}_h\right)$
   3. Update: $\theta_{t+1} = \theta_t + \eta_t \widetilde{\nabla}_\theta J(\theta_t)$

The Advantage Function (finite horizon)

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \middle| s_h = s\right] \qquad Q_h^\pi(s,a) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \middle| (s_h, a_h) = (s, a)\right]$$

The Advantage Function (finite horizon)

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \middle| s_h = s\right] \qquad Q_h^\pi(s,a) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \middle| (s_h, a_h) = (s,a)\right]$$

- The Advantage function is defined as:

$$A_h^\pi(s,a) = Q_h^\pi(s,a) - V_h^\pi(s,a)$$

The Advantage Function (finite horizon)

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \middle| s_h = s\right] \qquad Q_h^\pi(s,a) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \middle| (s_h, a_h) = (s,a)\right]$$

- The Advantage function is defined as:

$$A_h^\pi(s,a) = Q_h^\pi(s,a) - V_h^\pi(s,a)$$

- We have that:

$$E_{a\sim\pi(\cdot|s)}\left[A_h(s,a) \middle| s, h\right] = \sum_a \pi(a\,|\,s)A_h(s,a) = \;\; ??$$

The Advantage Function (finite horizon)

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \middle| s_h = s\right] \qquad Q_h^\pi(s,a) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \middle| (s_h, a_h) = (s,a)\right]$$

- The Advantage function is defined as:

$$A_h^\pi(s,a) = Q_h^\pi(s,a) - V_h^\pi(s,a)$$

- We have that:

$$E_{a\sim\pi(\cdot|s)}\left[A_h(s,a) \middle| s, h\right] = \sum_a \pi(a\,|\,s) A_h(s,a) = \ \ ??$$

- For the discounted case, $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s,a)$

The Advantage-based PG:

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

The Advantage-based PG:

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

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

- The second step follows by choosing $b_h(s) = V_h^\pi(s)$.
- In practice, the most common approach is to use $b_h(s)$ as an estimate of $V_h^\pi(s)$.

Summary so far:

Variance reduction with:

- Improvement over REINFORCE
- baseline functions (and the “advantage” formulation)

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

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

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

1-minute feedback form: https://bit.ly/3RHtlxy