Lecture 8:
Policy Gradient Algorithms

B. Ravindran

Solution Methods

- Value-based Methods
  - Q-learning
  - SARSA
  - TD(λ)
  - Actor-Critic

- Policy Search
  - Policy Gradient Methods
  - Evolutionary algorithms

- Model based methods
  - Stochastic Dynamic Programming
  - Bayesian approaches

Solution Methods

- Value-based Methods
  - Q-learning
  - SARSA
  - TD(λ)
  - Actor-Critic
- Policy Search
  - Policy Gradient Methods
  - ~~Evolutionary algorithms~~
- Model based methods
  - Stochastic Dynamic Programming
  - ~~Bayesian approaches~~

Policy Search Methods

- Policy search: Instead of maintaining estimates of value functions, search in the space of policies

- **Why?**
  - Simpler description
  - Better convergence
  - Robust to partial observability
  - Continuous action space

- Direct policy search - Genetic algorithms
- Policy Gradient Approaches

Policy Gradient Methods

- Policy depends on some parameters θ
  - Action preferences
  - Mean and variance
  - Weights of a neural network

- **Idea:** Modify policy parameters directly instead of estimating the action values

- Maximize: $J(\theta) = \mathbb{E}[G_t]$ $J(\theta) = \mathbb{E}[r_t]$

Simplified Setting

Immediate Reward or Multi-arm bandits

- θ update: $\theta \leftarrow \theta + \alpha \nabla J(\theta)$

Stochastic Gradient Ascent

- We compute the gradient of the performance *J(θ)* w.r.t the parameters *θ*

$$J(\theta) = E\left(r_t\right) = \sum_a q_*(a)\pi_\theta(a)$$

Bandit Setting
(Immediate Reward)

$$\nabla J(\theta) = \sum_a q_*(a)\nabla\pi_\theta(a)$$

$$= \sum_a q_*(a)\frac{\nabla\pi_\theta(a)}{\pi_\theta(a)}\pi_\theta(a) = \mathbb{E}_{\pi(\theta)}\left[q_*(a)\nabla\ln\pi_\theta(a)\right]$$

- Estimate the gradient from *N* samples

$$\hat{\nabla}J(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} r_i \underbrace{\frac{\nabla\pi_{\boldsymbol{\theta}}\left(a_i\right)}{\pi_{\boldsymbol{\theta}}\left(a_i\right)}}_{\substack{\text{Likelihood}\\ \text{Ratio}}}$$

REINFORCE

- REward Increment = Non-negative Factor × Offset Reinforcement × Characteristic Eligibility *(Williams '92)*

Bandit Setting
(Immediate Reward)

- Incremental version

$$\Delta\boldsymbol{\theta}_t = \alpha r_t \frac{\nabla \pi_{\boldsymbol{\theta}}(a_t)}{\pi_{\boldsymbol{\theta}}(a_t)}$$

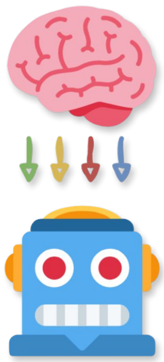
$$\Delta\boldsymbol{\theta}_t = \alpha r_t \frac{\partial \ln \pi_{\boldsymbol{\theta}}(a_t)}{\partial \boldsymbol{\theta}}$$

- REINFORCE with baseline

$$\Delta\boldsymbol{\theta}_t = \alpha(r_t - \boxed{b_t})\boxed{\frac{\partial \ln \pi_{\boldsymbol{\theta}}(a_t)}{\partial \boldsymbol{\theta}}}$$

Reinforcement
Baseline

Characteristic
Eligibility



Policy Gradient Theorem

Policy Gradient Theorem

- In the episodic case, we define the performance by assuming that every episode starts from state s_0 (non-random), as follows:

$$J(\boldsymbol{\theta}) \doteq v_{\pi_{\boldsymbol{\theta}}}(s_0)$$

  where $v_{\pi_{\boldsymbol{\theta}}}(s_0)$ is the true value function given a parameterized policy $\pi_{\boldsymbol{\theta}}$

- The Policy Gradient Theorem: The gradient of the performance can be expressed in terms of the gradient of the policy, as follows

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla \pi(a|s, \theta)$$

Policy Gradient Theorem

- Provides an analytic expression for the gradient of performance with respect to the policy parameter θ
- We begin the proof by expressing the gradient of the state-value function in terms of the action-value function

$$\nabla_\theta v_\pi(s) = \nabla_\theta \left(\sum_{a \in \mathcal{A}} \pi_\theta(a \mid s) q_\pi(s,a) \right)$$

Derivative product rule

$$= \sum_{a \in \mathcal{A}} \left(\nabla_\theta \pi_\theta(a \mid s) q_\pi(s,a) + \pi_\theta(a \mid s) \nabla_\theta q_\pi(s,a) \right)$$

Write Q_Π as V_Π

$$= \sum_{a \in \mathcal{A}} \left(\nabla_\theta \pi_\theta(a \mid s) q_\pi(s,a) + \pi_\theta(a \mid s) \nabla_\theta \sum_{s',r} p\left(s', r \mid s, a\right) \left(r + v_\pi\left(s'\right)\right) \right)$$

P is not a function of θ

$$= \sum_{a \in \mathcal{A}} \left(\nabla_\theta \pi_\theta(a \mid s) q_\pi(s,a) + \pi_\theta(a \mid s) \sum_{s',r} p\left(s', r \mid s, a\right) \nabla_\theta v_\pi\left(s'\right) \right)$$

Σ_r $P(r|s,a) = 1$

$$= \sum_{a \in \mathcal{A}} \left(\nabla_\theta \pi_\theta(a \mid s) q_\pi(s,a) + \pi_\theta(a \mid s) \sum_{s'} p\left(s' \mid s, a\right) \nabla_\theta v_\pi\left(s'\right) \right)$$

Policy Gradient Theorem

$$\nabla_\theta v_\pi(s) = \sum_{a \in \mathcal{A}} \left(\nabla_\theta \pi_\theta(a \mid s) q_\pi(s,a) + \pi_\theta(a \mid s) \sum_{s'} p\left(s' \mid s, a\right) \nabla_\theta v_\pi\left(s'\right) \right)$$

Unroll $\nabla_\theta V_\Pi(s')$

$$\nabla_\theta v_\pi(s) = \sum_{a \in \mathcal{A}} \left(\nabla_\theta \pi_\theta(a \mid s) q_\pi(s,a) + \pi_\theta(a \mid s) \sum_{s'} p\left(s' \mid s, a\right) \right.$$

$$\left. \sum_{a' \in \mathcal{A}} \left(\nabla_\theta \pi_\theta\left(a' \mid s'\right) q_\pi\left(s', a'\right) + \pi_\theta\left(a' \mid s'\right) \sum_{s''} p\left(s'' \mid s', a'\right) \nabla_\theta v_\pi\left(s''\right) \right) \right)$$

Policy Gradient Theorem

$$s \xrightarrow{a \sim \pi_\theta(.|s)} s' \xrightarrow{a \sim \pi_\theta(.|s')} s'' \xrightarrow{a \sim \pi_\theta(.|s'')} \ldots$$

Probability of transitioning from s to s in 0 steps while following π_θ

$$\rho_\pi(s \rightarrow s, k=0) = 1$$

Probability of transitioning from s to s' in 1 step while following π_θ

$$\rho_\pi(s \rightarrow s', k=1) = \sum_a \pi_\theta(a|s) p(s'|s,a)$$

Probability of transitioning from s to s' in in k steps and from s' to x in 1 step while following π_θ

$$\rho_\pi(s \rightarrow x, k+1) = \sum_{s'} \rho_\pi(s \rightarrow s', k) \rho_\pi(s' \rightarrow x, 1)$$

Policy Gradient Theorem

$$\nabla_\theta v_\pi(s) = \sum_{a\in\mathcal{A}}\left(\nabla_\theta \pi_\theta(a\mid s) q_\pi(s,a) + \pi_\theta(a\mid s)\sum_{s'} p\left(s'\mid s,a\right)\nabla_\theta v_\pi\left(s'\right)\right)$$

Unroll ∇θVπ(s')

$$\nabla_\theta v_\pi(s) = \sum_{a\in\mathcal{A}}\left(\nabla_\theta \pi_\theta(a\mid s) q_\pi(s,a) + \pi_\theta(a\mid s)\sum_{s'} p\left(s'\mid s,a\right)\right.$$

$$\left.\sum_{a'\in\mathcal{A}}\left(\nabla_\theta \pi_\theta\left(a'\mid s'\right) q_\pi\left(s',a'\right) + \pi_\theta\left(a'\mid s'\right)\sum_{s''} p\left(s''\mid s',a'\right)\nabla_\theta v_\pi\left(s''\right)\right)\right)$$

$$\phi(s) = \textstyle\sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a|s) q_\pi(s,a)$$

Distribute outer-most Σa

$$\nabla_\theta v_\pi(s) = \left(\underbrace{\sum_{a\in\mathcal{A}}\nabla_\theta \pi_\theta(a\mid s) q_\pi(s,a)}_{\phi(s)} + \sum_{s'}\underbrace{\sum_{a\in\mathcal{A}}\pi_\theta(a\mid s) p\left(s'\mid s,a\right)}_{\sum_{s'}\rho_\pi(s\to s',k=1)}\right.$$

$$\left.\underbrace{\sum_{a'\in\mathcal{A}}\left(\nabla_\theta \pi_\theta\left(a'\mid s'\right) q_\pi\left(s',a'\right)\right.}_{\phi(s')} + \pi_\theta\left(a'\mid s'\right)\sum_{s''} p\left(s''\mid s',a'\right)\nabla_\theta v_\pi\left(s''\right)\right)\Bigg)$$

Policy Gradient Theorem

$$\nabla_\theta v_\pi(s) = \Bigg(\underbrace{\sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a \mid s) q_\pi(s,a)}_{\phi(s)} + \sum_{s'} \underbrace{\sum_{a\in\mathcal{A}} \pi_\theta(a \mid s) p\left(s' \mid s, a\right)}_{\sum_{s'} \rho_\pi(s \to s', k=1)}$$

$$\underbrace{\sum_{a'\in\mathcal{A}} \Bigg(\nabla_\theta \pi_\theta\left(a' \mid s'\right) q_\pi\left(s', a'\right)}_{\phi(s')} + \pi_\theta\left(a' \mid s'\right) \sum_{s''} p\left(s'' \mid s', a'\right) \nabla_\theta v_\pi\left(s''\right)\Bigg)\Bigg)$$

$$\nabla_\theta v_\pi(s) = \phi(s) + \sum_{s'} \rho_\pi(s \to s', 1)[\phi(s') + \sum_{s''} \rho_\pi(s' \to s'', 1) \nabla_\theta v_\pi\left(s''\right)]$$

Distribute $\Sigma_{s'}$

$$\nabla_\theta v_\pi(s) = \phi(s) + \sum_{s'} \rho_\pi(s \to s', 1)\phi(s') + \sum_{s''} \rho_\pi(s' \to s'', 2) \nabla_\theta v_\pi\left(s''\right)$$

Repeatedly unroll till ∞

$$\nabla_\theta v_\pi(s) = \sum_{x\in\mathcal{S}} \sum_{k=0}^{\infty} \rho_\pi(s \to x, k)\phi(x)$$

Policy Gradient Theorem

$$\nabla J(\theta) = \nabla_\theta v_\pi(s_0)$$ *(Remember)*

$$\nabla_\theta v_\pi(s_0) = \sum_{s\in\mathcal{S}} \left(\sum_{k=0}^{\infty} \rho_\pi(s_0 \to s, k) \right) \sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a|s) q_\pi(s,a)$$

$$\nabla_\theta v_\pi(s_0) = \sum_{s\in\mathcal{S}} \eta(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a|s) q_\pi(s,a)$$

Introduce the probability version of η(s)

$$\nabla_\theta v_\pi(s_0) = \sum_{s'\in\mathcal{S}} \eta(s') \sum_{s\in\mathcal{S}} \frac{\eta(s)}{\sum_{x\in\mathcal{S}} \eta(x)} \sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a|s) q_\pi(s,a)$$

$$\nabla_\theta v_\pi(s_0) = \sum_{s'\in\mathcal{S}} \eta(s') \sum_{s\in\mathcal{S}} \mu(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a|s) q_\pi(s,a)$$

$$\boxed{\nabla_\theta v_\pi(s_0) \propto \sum_{s\in\mathcal{S}} \mu(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a|s) q_\pi(s,a)}$$

η(s') is a constant,
= 1 (in continuing task)
= avg episodic length (o.w)

REINFORCE: MC Policy Gradient

$$
\begin{aligned}
\nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\boldsymbol{\theta}) \\
&= \mathbb{E}_\pi\left[\sum_a q_\pi(S_t,a) \nabla \pi(a|S_t,\boldsymbol{\theta})\right] \\
&= \mathbb{E}_\pi\left[q_\pi(S_t,A_t) \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t,\boldsymbol{\theta})}\right] \\
&= \mathbb{E}_\pi\left[G_t \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta})}{\pi(A_t|S_t,\boldsymbol{\theta})}\right]
\end{aligned}
$$

Expectation over the visited states while following Π

Multiply and divide by Π(a|s,θ) +

Expectation over the actions taken while following Π

$$
\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta}_t)}{\pi(A_t|S_t,\boldsymbol{\theta}_t)}
$$

The gradient update is

- ❏ proportional to the return
- ❏ inversely proportional to the action probability

REINFORCE: MC PG Control

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Algorithm parameter: step size $\alpha > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \gamma^t G \nabla \ln \pi(A_t|S_t, \boldsymbol{\theta})$

- The algorithm computes an unbiased estimate of the gradient
- Can be very slow due to high variance in the estimates
- Variance is related to the “recurrence time” or the episode length
- In large state spaces, the variance becomes unacceptably high

REINFORCE w/ Baseline

- The policy gradient theorem can be generalized to include a comparison of the action value to an arbitrary baseline *b(s)*

$$\nabla J(\boldsymbol{\theta}) \propto \sum_s \mu(s) \sum_a \Big(q_\pi(s,a) - b(s) \Big) \nabla \pi(a|s,\boldsymbol{\theta})$$

- Baseline - should be a function that is independent on the action *a*

$$\sum_a b(s) \nabla \pi(a|s,\boldsymbol{\theta}) \;=\; b(s) \nabla \sum_a \pi(a|s,\boldsymbol{\theta}) \;=\; b(s) \nabla 1 \;=\; 0$$

- Update rule of REINFORCE with baseline

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha \Big(G_t - b(S_t) \Big) \frac{\nabla \pi(A_t|S_t,\boldsymbol{\theta}_t)}{\pi(A_t|S_t,\boldsymbol{\theta}_t)}$$

(Doesn’t change the expected value, but has an effect on the variance)

Actor-Critic Methods

- Actor-Critic methods learn both a policy and a state-value function simultaneously
- The policy is referred to as the actor that suggests actions given a state
- The value function is referred to as the critic. It evaluates actions taken by the actor based on the given policy

$$
\begin{aligned}
\boldsymbol{\theta}_{t+1} &\doteq \boldsymbol{\theta}_t + \alpha\left(G_t - \hat{v}\left(S_t, \mathbf{w}\right)\right) \frac{\nabla \pi\left(A_t \mid S_t, \boldsymbol{\theta}_t\right)}{\pi\left(A_t \mid S_t, \boldsymbol{\theta}_t\right)} \\
&= \boldsymbol{\theta}_t + \alpha\left(R_{t+1} + \gamma \hat{v}\left(S_{t+1}, \mathbf{w}\right) - \hat{v}\left(S_t, \mathbf{w}\right)\right) \frac{\nabla \pi\left(A_t \mid S_t, \boldsymbol{\theta}_t\right)}{\pi\left(A_t \mid S_t, \boldsymbol{\theta}_t\right)} \\
&= \boldsymbol{\theta}_t + \alpha\, \delta_t \frac{\nabla \pi\left(A_t \mid S_t, \boldsymbol{\theta}_t\right)}{\pi\left(A_t \mid S_t, \boldsymbol{\theta}_t\right)}
\end{aligned}
$$

One-step Actor-Critic Algorithm

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)
Loop forever (for each episode):
    Initialize S (first state of episode)
    $I \leftarrow 1$
    Loop while S is not terminal (for each time step):
        $A \sim \pi(\cdot|S, \boldsymbol{\theta})$
        Take action A, observe S', R
        $\delta \leftarrow R + \gamma \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$ (if S' is terminal, then $\hat{v}(S',\mathbf{w}) \doteq 0$)
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S,\mathbf{w})$
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} I \delta \nabla \ln \pi(A|S, \boldsymbol{\theta})$
        $I \leftarrow \gamma I$
        $S \leftarrow S'$

(This is a fully online, incremental algorithm, with states, actions, and rewards processed as they occur and then never revisited again)