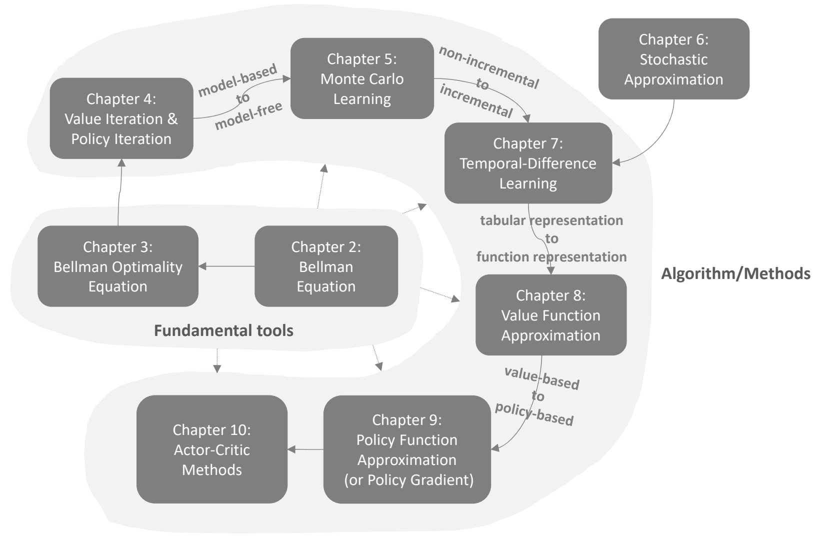
Lecture 9: Policy Gradient Methods

Shiyu Zhao

Introduction



Introduction

In this lecture, we will move

- from value-based methods to policy-based methods
- from value function approximation to policy function approximation

Outline

Outline

Basic idea of policy gradient

Previously, policies have been represented by tables:

- The action probabilities of all states are stored in a table $\pi(a|s)$. Each entry of the table is indexed by a state and an action.

	a_1	a_2	a_3	a_4	a_5
s_1	$\pi(a_1\|s_1)$	$\pi(a_2\|s_1)$	$\pi(a_3\|s_1)$	$\pi(a_4\|s_1)$	$\pi(a_5\|s_1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s_9	$\pi(a_1\|s_9)$	$\pi(a_2\|s_9)$	$\pi(a_3\|s_9)$	$\pi(a_4\|s_9)$	$\pi(a_5\|s_9)$

- We can directly access or change a value in the table.

Basic idea of policy gradient

Now, policies can be represented by parameterized functions:

$$\pi(a|s, \theta)$$

where $\theta \in \mathbb{R}^m$ is a parameter vector.

- The function can be, for example, a neural network, whose input is s, output is the probability to take each action, and parameter is θ.
- **Advantage:** when the state space is large, the tabular representation will be of low efficiency in terms of storage and generalization.
- The function representation is also sometimes written as $\pi(a, s, \theta)$, $\pi_\theta(a|s)$, or $\pi_\theta(a, s)$.

Basic idea of policy gradient

Differences between tabular and function representations:

- First, how to define optimal policies?
  - When represented as a table, a policy π is optimal if it can maximize *every state value*.
  - When represented by a function, a policy π is optimal if it can maximize certain *scalar metrics*.

Differences between tabular and function representations:

- Second, how to access the probability of an action?
  - In the tabular case, the probability of taking a at s can be directly accessed by looking up the tabular policy.
  - In the case of function representation, we need to calculate the value of $\pi(a|s, \theta)$ given the function structure and the parameter.

Basic idea of policy gradient

Differences between tabular and function representations:

- Third, how to update policies?
  - When represented by a table, a policy π can be updated by directly changing the entries in the table.
  - When represented by a parameterized function, a policy π cannot be updated in this way anymore. Instead, it can only be updated by changing *the parameter* θ.

Basic idea of policy gradient

The basic idea of the policy gradient is simple:

- First, metrics (or objective functions) to define optimal policies: $J(\theta)$, which can define optimal policies.
- Second, gradient-based optimization algorithms to search for optimal policies:

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\theta_t)$$

Although the idea is simple, the complication emerges when we try to answer the following questions.

- What appropriate metrics should be used?
- How to calculate the gradients of the metrics?

These questions will be answered in detail in this lecture.

Outline

Metrics to define optimal policies - 1) The average value

There are two metrics.

The first metric is the average state value or simply called average value. In particular, the metric is defined as

$$\bar{v}_\pi = \sum_{s \in \mathcal{S}} d(s) v_\pi(s)$$

- $\bar{v}_\pi$ is a weighted average of the state values.
- $d(s) \geq 0$ is the weight for state s.
- Since $\sum_{s \in \mathcal{S}} d(s) = 1$, we can interpret $d(s)$ as a probability distribution. Then, the metric can be written as

$$\bar{v}_\pi = \mathbb{E}[v_\pi(S)]$$

where $S \sim d$.

Metrics to define optimal policies - 1) The average value

Vector-product form:

$$\bar{v}_\pi = \sum_{s \in \mathcal{S}} d(s) v_\pi(s) = d^T v_\pi$$

where

$$v_\pi = [\ldots, v_\pi(s), \ldots]^T \in \mathbb{R}^{|\mathcal{S}|}$$
$$d = [\ldots, d(s), \ldots]^T \in \mathbb{R}^{|\mathcal{S}|}.$$

This expression is particularly useful when we analyze its gradient.

Metrics to define optimal policies - 1) The average value

How to select the distribution d? There are two cases.

The first case is that d is **independent** of the policy π.

- This case is relatively simple because the gradient of the metric is easier to calculate.
- In this case, we specifically denote d as d_0 and $\bar{v}_\pi$ as $\bar{v}_\pi^0$.
- How to select d_0?
  - One trivial way is to treat all the states equally important and hence select $d_0(s) = 1/|\mathcal{S}|$.
  - Another important case is that we are only interested in a specific state s_0. For example, the episodes in some tasks always start from the same state s_0. Then, we only care about the long-term return starting from s_0. In this case,

$$d_0(s_0) = 1, \quad d_0(s \neq s_0) = 0.$$

Metrics to define optimal policies - 1) The average value

How to select the distribution d? There are two cases.

The second case is that d **depends** on the policy π.

- A common way to select d as $d_\pi(s)$, which is the stationary distribution under π. Details of stationary distribution can be found in the last lecture and the book.
  - One basic property of d_π is that it satisfies

    $$d_\pi^T P_\pi = d_\pi^T,$$

    where P_π is the state transition probability matrix.
  - The interpretation of selecting d_π is as follows.
    - If one state is frequently visited in the long run, it is more important and deserves more weight.
    - If a state is hardly visited, then we give it less weight.

Metrics to define optimal policies - 2) The average reward

The second metric is average one-step reward or simply average reward. In particular, the metric is

$$\bar{r}_\pi \doteq \sum_{s \in \mathcal{S}} d_\pi(s) r_\pi(s) = \mathbb{E}[r_\pi(S)],$$

where $S \sim d_\pi$. Here,

$$r_\pi(s) \doteq \sum_{a \in \mathcal{A}} \pi(a|s) r(s,a)$$

is the average of the one-step immediate reward that can be obtained starting from state s, and

$$r(s,a) = \mathbb{E}[R|s,a] = \sum_r r p(r|s,a)$$

- The weight d_π is the stationary distribution.
- As its name suggests, $\bar{r}_\pi$ is simply a weighted average of the one-step immediate rewards.

Metrics to define optimal policies - 2) The average reward

An equivalent definition!

- Suppose an agent follows a given policy and generate a trajectory with the rewards as $(R_{t+1}, R_{t+2}, \dots)$.
- The average single-step reward along this trajectory is

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\Big[R_{t+1} + R_{t+2} + \cdots + R_{t+n} | S_t = s_0\Big] \\ =&\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{k=1}^{n} R_{t+k} | S_t = s_0\right] \end{aligned}$$

where s_0 is the starting state of the trajectory.

Metrics to define optimal policies - Remarks

An important property is that

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\sum_{k=1}^{n}R_{t+k}|S_t=s_0\right] &= \lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\sum_{k=1}^{n}R_{t+k}\right]\\ &= \sum_s d_\pi(s)r_\pi(s)\\ &= \bar{r}_\pi\end{aligned}$$

Note that

- The starting state s_0 does not matter.
- The two definitions of $\bar{r}_\pi$ are equivalent.

See the proof in the book.

Metrics to define optimal policies - Remarks

Remark 1 about the metrics:

- All these metrics are functions of π.
- Since π is parameterized by θ, these metrics are functions of θ.
- In other words, different values of θ can generate different metric values.
- Therefore, we can search for the optimal values of θ to maximize these metrics.

This is the basic idea of policy gradient methods.

Remark 2 about the metrics:

- One complication is that the metrics can be defined in either the discounted case where $\gamma \in (0, 1)$ or the undiscounted case where $\gamma = 1$.
- We only consider the discounted case so far in this book. For details about the undiscounted case, see the book.

Metrics to define optimal policies - Remarks

Remark 3 about the metrics:

- Intuitively, $\bar{r}_\pi$ is more short-sighted because it merely considers the immediate rewards, whereas $\bar{v}_\pi$ considers the total reward overall steps.
- However, the two metrics are equivalent to each other.
  In the discounted case where $\gamma < 1$, it holds that

  $$\bar{r}_\pi = (1-\gamma)\bar{v}_\pi.$$

  See the proof in the book.

Metrics to define optimal policies - Excise

Excise:

You will see the following metric often in the literature:

$$J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\right]$$

What is its relationship to the metrics we introduced just now?

Metrics to define optimal policies - Excise

$$J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\right]$$

Answer: First, clarify and understand this metric.

- It starts from $S_0 \sim d$ and then $A_0, R_1, S_1, A_1, R_2, S_2, \dots$
- $A_t \sim \pi(S_t)$ and $R_{t+1}, S_{t+1} \sim p(R_{t+1}|S_t, A_t), p(S_{t+1}|S_t, A_t)$

Then, we know this metric is the same as the average value because

$$\begin{aligned} J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\right] &= \sum_{s \in \mathcal{S}} d(s) \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | S_0 = s\right] \\ &= \sum_{s \in \mathcal{S}} d(s) v_\pi(s) \\ &= \bar{v}_\pi \end{aligned}$$

Outline

Gradients of the metrics

Given a metric, we next

- derive its gradient
- and then, apply gradient-based methods to optimize the metric.

The gradient calculation is one of the most complicated parts of policy gradient methods! That is because

- first, we need to distinguish different metrics $\bar{v}_\pi$, $\bar{r}_\pi$, $\bar{v}_\pi^0$
- second, we need to distinguish the discounted and undiscounted cases.

The calculation of the gradients:

- We will not discuss the details in this lecture.
- Interested readers may see my book for details.

Gradients of the metrics

Summary of the results about the gradients:

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

where

- $J(\theta)$ can be $\bar{v}_\pi$, $\bar{r}_\pi$, or $\bar{v}_\pi^0$.
- "$=$" may denote strict equality, approximation, or proportional to.
- η is a distribution or weight of the states.

Gradients of the metrics

Some specific results:

$$\nabla_\theta \bar{r}_\pi \simeq \sum_s d_\pi(s) \sum_a \nabla_\theta \pi(a|s,\theta) q_\pi(s,a),$$

$$\nabla_\theta \bar{v}_\pi = \frac{1}{1-\gamma} \nabla_\theta \bar{r}_\pi$$

$$\nabla_\theta \bar{v}_\pi^0 = \sum_{s \in \mathcal{S}} \rho_\pi(s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a)$$

Details are not given here. Interested readers can read my book.

Gradients of the metrics

A compact and useful form of the gradient:

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{s\in\mathcal{S}} \eta(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a) \\
&= \mathbb{E}\big[\nabla_\theta \ln \pi(A|S,\theta) q_\pi(S,A)\big]
\end{aligned}
$$

where $S \sim \eta$ and $A \sim \pi(A|S,\theta)$.

Why is this expression useful?

- Because we can use samples to approximate the gradient!

$$
\nabla_\theta J \approx \nabla_\theta \ln \pi(a|s,\theta) q_\pi(s,a)
$$

Gradients of the metrics

$$\nabla_\theta J(\theta) = \sum_{s\in\mathcal{S}} \eta(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a)$$
$$= \mathbb{E}\big[\nabla_\theta \ln \pi(A|S,\theta) q_\pi(S,A)\big]$$

How to prove the above equation?

Consider the function $\ln \pi$ where $\ln$ is the natural logarithm. It is easy to see that

$$\nabla_\theta \ln \pi(a|s,\theta) = \frac{\nabla_\theta \pi(a|s,\theta)}{\pi(a|s,\theta)}$$

and hence

$$\nabla_\theta \pi(a|s,\theta) = \pi(a|s,\theta) \nabla_\theta \ln \pi(a|s,\theta).$$

Gradients of the metrics

Then, we have

$$
\begin{aligned}
\nabla_\theta J &= \sum_s d(s) \sum_a \nabla_\theta \pi(a|s,\theta) q_\pi(s,a) \\
&= \sum_s d(s) \sum_a \pi(a|s,\theta) \nabla_\theta \ln \pi(a|s,\theta) q_\pi(s,a) \\
&= \mathbb{E}_{S\sim d}\left[\sum_a \pi(a|S,\theta)\nabla_\theta \ln \pi(a|S,\theta) q_\pi(S,a)\right] \\
&= \mathbb{E}_{S\sim d, A\sim\pi}\big[\nabla_\theta \ln \pi(A|S,\theta) q_\pi(S,A)\big] \\
&\doteq \mathbb{E}\big[\nabla_\theta \ln \pi(A|S,\theta) q_\pi(S,A)\big]
\end{aligned}
$$

Gradients of the metrics

Some remarks: Because we need to calculate $\ln \pi(a|s,\theta)$, we must ensure that for all s, a, θ

$$\pi(a|s,\theta) > 0$$

- This can be archived by using softmax functions that can normalize the entries in a vector from $(-\infty, +\infty)$ to $(0, 1)$.
- For example, for any vector $x = [x_1, \dots, x_n]^T$,

$$z_i = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}$$

  where $z_i \in (0, 1)$ and $\sum_{i=1}^n z_i = 1$.
- Then, the policy function has the form of

$$\pi(a|s,\theta) = \frac{e^{h(s,a,\theta)}}{\sum_{a' \in \mathcal{A}} e^{h(s,a',\theta)}},$$

  where $h(s, a, \theta)$ is another function.

Gradients of the metrics

Some remarks:

- Such a form based on the softmax function can be realized by a neural network whose input is s and parameter is θ. The network has $|\mathcal{A}|$ outputs, each of which corresponds to $\pi(a|s,\theta)$ for an action a. The activation function of the output layer should be softmax.
- Since $\pi(a|s,\theta) > 0$ for all a, the parameterized policy is stochastic and hence exploratory.
- There also exist deterministic policy gradient (DPG) methods.

Outline

Gradient-ascent algorithm

Now, we are ready to present the first policy gradient algorithm to find optimal policies!

- The gradient-ascent algorithm maximizing $J(\theta)$ is

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha \nabla_\theta J(\theta) \\ &= \theta_t + \alpha \mathbb{E}\Big[\nabla_\theta \ln \pi(A|S,\theta_t) q_\pi(S,A)\Big]\end{aligned}$$

- The true gradient can be replaced by a stochastic one:

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta \ln \pi(a_t|s_t,\theta_t) q_\pi(s_t,a_t)$$

Gradient-ascent algorithm

- Furthermore, since q_π is unknown, it can be approximated:

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta \ln \pi(a_t|s_t, \theta_t) q_t(s_t, a_t)$$

There are different methods to approximate $q_\pi(s_t, a_t)$

- In this lecture, Monte-Carlo based method, *REINFORCE*
- In the next lecture, TD method and more

Remark 1: How to do sampling?

$$\mathbb{E}_{S\sim d, A\sim\pi}\Big[\nabla_\theta \ln \pi(A|S,\theta_t) q_\pi(S,A)\Big] \longrightarrow \nabla_\theta \ln \pi(a|s,\theta_t) q_\pi(s,a)$$

- How to sample S?
  - $S \sim d$, where the distribution d is a long-run behavior under π.
- How to sample A?
  - $A \sim \pi(A|S,\theta)$. Hence, a_t should be sampled following $\pi(\theta_t)$ at s_t.
  - Therefore, the policy gradient method is on-policy.

Remark 2: How to interpret this algorithm?

Since

$$\nabla_\theta \ln \pi(a_t|s_t,\theta_t) = \frac{\nabla_\theta \pi(a_t|s_t,\theta_t)}{\pi(a_t|s_t,\theta_t)}$$

the algorithm can be rewritten as

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha \nabla_\theta \ln \pi(a_t|s_t,\theta_t) q_t(s_t,a_t) \\ &= \theta_t + \alpha \underbrace{\left(\frac{q_t(s_t,a_t)}{\pi(a_t|s_t,\theta_t)}\right)}_{\beta_t} \nabla_\theta \pi(a_t|s_t,\theta_t).\end{aligned}$$

Therefore, we have the important expression of the algorithm:

$$\theta_{t+1} = \theta_t + \alpha \beta_t \nabla_\theta \pi(a_t|s_t,\theta_t)$$

Gradient-ascent algorithm

It is a gradient-ascent algorithm for maximizing $\pi(a_t|s_t,\theta)$:

$$\theta_{t+1} = \theta_t + \alpha\beta_t \nabla_\theta \pi(a_t|s_t,\theta_t)$$

Intuition: When $\alpha\beta_t$ is sufficiently small

- If $\beta_t > 0$, the probability of choosing (s_t, a_t) is enhanced:

  $$\pi(a_t|s_t,\theta_{t+1}) > \pi(a_t|s_t,\theta_t)$$

  The greater β_t is, the stronger the enhancement is.
- If $\beta_t < 0$, then $\pi(a_t|s_t,\theta_{t+1}) < \pi(a_t|s_t,\theta_t)$.

Math: When $\theta_{t+1} - \theta_t$ is sufficiently small, we have

$$\begin{aligned}
\pi(a_t|s_t,\theta_{t+1}) &\approx \pi(a_t|s_t,\theta_t) + (\nabla_\theta \pi(a_t|s_t,\theta_t))^T(\theta_{t+1} - \theta_t) \\
&= \pi(a_t|s_t,\theta_t) + \alpha\beta_t (\nabla_\theta \pi(a_t|s_t,\theta_t))^T(\nabla_\theta \pi(a_t|s_t,\theta_t)) \\
&= \pi(a_t|s_t,\theta_t) + \alpha\beta_t \|\nabla_\theta \pi(a_t|s_t,\theta_t)\|^2
\end{aligned}$$

Gradient-ascent algorithm

$$\theta_{t+1} = \theta_t + \alpha \underbrace{\left(\frac{q_t(s_t, a_t)}{\pi(a_t|s_t, \theta_t)} \right)}_{\beta_t} \nabla_\theta \pi(a_t|s_t, \theta_t)$$

The coefficient β_t can well balance exploration and exploitation.

- First, β_t is proportional to $q_t(s_t, a_t)$.
  - If $q_t(s_t, a_t)$ is great, then β_t is great.
  - Therefore, the algorithm intends to enhance actions with greater values.
- Second, β_t is inversely proportional to $\pi(a_t|s_t, \theta_t)$.
  - If $\pi(a_t|s_t, \theta_t)$ is small, then β_t is large.
  - Therefore, the algorithm intends to explore actions that have low probabilities.

REINFORCE algorithm

Recall that

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta \ln \pi(a_t|s_t, \theta_t) q_\pi(s_t, a_t)$$

is replaced by

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta \ln \pi(a_t|s_t, \theta_t) q_t(s_t, a_t)$$

where $q_t(s_t, a_t)$ is an approximation of $q_\pi(s_t, a_t)$.

- If $q_\pi(s_t, a_t)$ is approximated by Monte Carlo estimation, the algorithm has a specifics name, REINFORCE.
- REINFORCE is one of earliest and simplest policy gradient algorithms.
- Many other policy gradient algorithms such as the actor-critic methods can be obtained by extending REINFORCE (next lecture).

REINFORCE algorithm

Pseudocode: Policy Gradient by Monte Carlo (REINFORCE)

Initialization: A parameterized function $\pi(a|s,\theta)$, $\gamma \in (0,1)$, and $\alpha > 0$.

Aim: Search for an optimal policy maximizing $J(\theta)$.

For the kth iteration, do

Select s_0 and generate an episode following $\pi(\theta_k)$. Suppose the episode is $\{s_0, a_0, r_1, \ldots, s_{T-1}, a_{T-1}, r_T\}$.

For $t = 0, 1, \ldots, T-1$, do

Value update: $q_t(s_t, a_t) = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k$

Policy update: $\theta_{t+1} = \theta_t + \alpha \nabla_\theta \ln \pi(a_t|s_t, \theta_t) q_t(s_t, a_t)$

$\theta_k = \theta_T$

Outline

Summary

Contents of this lecture:

- Metrics for optimality
- Gradients of the metrics
- Gradient-ascent algorithm
- A special case: REINFORCE

Next lecture: Actor-critic