

CS/Stat 184

# Introduction to Reinforcement Learning

# Contents

Welcome to the study of reinforcement learning! This set of lecture notes accompanies the undergraduate course CS/STAT 184 and is intended to be a friendly yet rigorous introduction to this exciting and active subfield of machine learning. Here are some questions you might have before embarking on this journey:

**What is reinforcement learning (RL)?** Broadly speaking, RL is a subfield of machine learning that studies how an agent can learn to make sequential decisions in an environment.

**Why study RL?** RL provides a powerful framework for attacking a wide variety of problems, including robotic control, video games and board games, resource management, language modelling, and more. It also provides an interdisciplinary paradigm for studying animal and human behavior. Many of the most stunning results in machine learning, ranging from AlphaGo to ChatGPT, are built on top of RL.

**Is this book for me?** This book assumes familiarity with multivariable calculus, linear algebra, and probability. For Harvard undergraduates, this would be fulfilled by Math 21a, Math 21b, and Stat 110. Stat 111 is strongly recommended but not required. Here is a non-comprehensive list of topics of which this book will assume knowledge:

- **Linear Algebra:** Vectors, matrices, matrix multiplication, matrix inversion, eigenvalues and eigenvectors, and the Gram-Schmidt process.
- **Multivariable Calculus:** Partial derivatives, gradient, directional derivative, and the chain rule.
- **Probability:** Random variables, probability distributions, expectation, variance, covariance, conditional probability, Bayes' rule, and the law of total probability.

**How does reinforcement learning differ from other machine learning paradigms?** Here is a list of comparisons:

- **Supervised learning.** Supervised learning concerns itself with learning a mapping from inputs to outputs (e.g. image classification). Typically the data takes the form of input-output pairs that are assumed to be sampled independently from some generating distribution. In RL, however, the data is generated by the agent interacting with the environment, meaning the observations depend on the agent's behaviour and are not independent from each other. This requires a more general set of tools.

Conversely, supervised learning is a well-studied field that provides many useful tools for RL. For example, it may be useful to use supervised learning to predict how valuable a given state is, or to predict the probability of transitioning to a given state.

## 0.1 Overview

Chapter 1 introduces **Markov Decision Processes**, the dominant mathematical framework for studying RL. We'll discuss **dynamic programming** algorithms for solving MDPs, including

**policy evaluation, policy iteration, and value iteration.**

Chapter 2 then discusses **multi-armed bandits**, a simpler problem that is often used as a warm-up to RL.

## 0.2 Notation

We will use the following notation throughout the book. This notation is inspired by Sutton and Barto and AJKS .

$s$	A state.
$a$	An action.
$r$	A reward.
$p$	A probability.
$\pi$	A policy.
$V$	A value function.
$Q$	An action-value function.
$A$	An advantage function.
$\gamma$	A discount factor.
$\tau$	A trajectory.
$\mathcal{S}$	A state space.
$\mathcal{A}$	An action space.

## 0.3 Challenges of reinforcement learning

**Exploration-exploitation tradeoff.** Should the agent try a new action or stick with the action that it knows is good?

**Prediction.** The agent might want to predict the value of a state or state-action pair.

**Policy computation (control).** In a complex environment, even if the dynamics are known, it can still be challenging to compute the best policy.

## 0.4 Resources

Inspired by the Stat 110 textbook and Stat 111 lecture notes.

This book seeks to provide an intuitive understanding before technical treatment.

# Contents

# Chapter 1

## Bandits

The **multi-armed bandits** (MAB) setting is a simple but powerful setting for studying the basic challenges of RL. In this setting, an agent repeatedly chooses from a fixed set of actions, called **arms**, each of which has an associated reward distribution. The agent's goal is to maximize the total reward it receives over some time period.

States	Actions	Rewards
None	Finite	Stochastic

In particular, we'll spend a lot of time discussing the **Exploration-Exploitation Tradeoff**: should the agent choose new actions to learn more about the environment, or should it choose actions that it already knows to be good?

### Example 1.0.1: Online advertising

Let's suppose you, the agent, are an advertising company. You have  $K$  different ads that you can show to users; For concreteness, let's suppose there's just a single user. You receive 1 reward if the user clicks the ad, and 0 otherwise. Thus, the unknown *reward distribution* associated to each ad is a Bernoulli distribution defined by the probability that the user clicks on the ad. Your goal is to maximize the total number of clicks by the user.

### Example 1.0.2: Clinical trials

Suppose you're a pharmaceutical company, and you're testing a new drug. You have  $K$  different dosages of the drug that you can administer to patients. You receive 1 reward if the patient recovers, and 0 otherwise. Thus, the unknown *reward distribution* associated to each dosage is a Bernoulli distribution defined by the probability that the patient recovers. Your goal is to maximize the total number of patients that recover.

In this chapter, we will introduce the multi-armed bandits setting, and discuss some of the challenges that arise when trying to solve problems in this setting. We will also introduce some of the key concepts that we will use throughout the book, such as regret and exploration-exploitation tradeoffs.

## 1.1 Introduction

### Remark 1.1.1: Namesake

The name “multi-armed bandits” comes from slot machines in casinos, which are often called “one-armed bandits” since they have one arm (the lever) and take money from the player.

Let  $K$  denote the number of arms. We’ll label them  $0, \dots, K-1$  and use *superscripts* to indicate the arm index; since we seldom need to raise a number to a power, this hopefully won’t cause much confusion. For simplicity, we’ll assume rewards are *bounded* between 0 and 1. Then each arm has an unknown reward distribution  $\nu^k \in \Delta([0, 1])$  with mean  $\mu^k = \mathbb{E}_{r \sim \nu^k}[r]$ .

In pseudocode, the agent’s interaction with the MAB environment can be described by the following process:

```

for  $t = 0, \dots, T$  do
    Agent chooses  $a_t \in [K]$ 
    Agent receives  $r_t \sim \nu^{a_t}$ 
    Agent updates its internal state
end for

```

What’s the *optimal* strategy for the agent, i.e. the one that achieves the highest expected reward? Convince yourself that the agent should try to always pull the arm with the highest expected reward  $\mu^* := \max_{k \in [K]} \mu^k$ .

The goal, then, can be rephrased as to minimize the **regret**, defined below:

### Definition 1.1.1: Regret

The agent’s **regret** after  $T$  timesteps is defined as

$$\text{Regret}_T := \sum_{t=0}^{T-1} \mu^* - \mu^{a_t} \quad (1.1)$$

Note that this depends on the *true means* of the pulled arms, *not* the actual observed rewards. We typically think of this as a random variable where the randomness comes from the agent’s strategy (i.e. the sequence of actions  $a_0, \dots, a_{T-1}$ ).

Throughout the chapter, we will try to upper bound the regret of various algorithms in two different senses:

1. Upper bound the *expected* regret, i.e. show  $\mathbb{E}[\text{Regret}_T] \leq M_T$ .
2. Find a high-probability upper bound on the regret, i.e. show  $\mathbb{P}(\text{Regret}_T \leq M_{T,\delta}) \geq 1 - \delta$ .

Note that these two different approaches say very different things about the regret. The first approach says that the *average* regret is at most  $M_T$ . However, the agent might still

achieve higher regret on many runs. The second approach says that, *with high probability*, the agent will achieve regret at most  $M_{T,\delta}$ . However, it doesn't say anything about the regret in the remaining  $\delta$  fraction of runs, which might be arbitrarily high.

We'd like to achieve **sublinear regret** in expectation, i.e.  $\mathbb{E}[\text{Regret}_T] = o(T)$ . That is, as we learn more about the environment, we'd like to be able to exploit that knowledge to achieve higher rewards.

The rest of the chapter comprises a series of increasingly sophisticated MAB algorithms.

## 1.2 Pure exploration (random guessing)

A trivial strategy is to always choose arms at random (i.e. “pure exploration”).

### Definition 1.2.1: Pure exploration

```

for  $t \leftarrow 0$  to  $T - 1$  do
  Choose  $a_t \sim \text{Unif}([K])$ 
  Observe  $r_t \sim \nu^{a_t}$ 
end for

```

Note that

$$\mathbb{E}_{a_t \sim \text{Unif}([K])}[\mu^{a_t}] = \bar{\mu} = \frac{1}{K} \sum_{k=1}^K \mu^k$$

so the expected regret is simply

$$\begin{aligned} \mathbb{E}[\text{Regret}_T] &= \sum_{t=0}^{T-1} \mathbb{E}[\mu^* - \mu^{a_t}] \\ &= T(\mu^* - \bar{\mu}) > 0. \end{aligned}$$

This scales as  $\Theta(T)$ , i.e. *linear* in the number of timesteps  $T$ . There's no learning here: the agent doesn't use any information about the environment to improve its strategy.

## 1.3 Pure greedy

How might we improve on pure exploration? Instead, we could try each arm once, and then commit to the one with the highest observed reward. We'll call this the **pure greedy** strategy.

### Definition 1.3.1: Pure greedy

```

for  $k \leftarrow 0$  to  $K - 1$  do
  Observe  $r^k \sim \nu^k$ 
end for

```

▷ Exploration phase



```

 $\hat{k} \leftarrow \arg \max_{k \in [K]} r^k$ 
for  $t \leftarrow K$  to  $T - 1$  do                                     ▷ Exploitation phase
    Observe  $r_t \sim \nu^{\hat{k}}$ 
end for

```

Note we've used superscripts  $r^k$  during the exploration phase to indicate that we observe exactly one reward for each arm. Then we use subscripts  $r_t$  during the exploitation phase to indicate that we observe a sequence of rewards from the chosen greedy arm  $\hat{k}$ .

How does the expected regret of this strategy compare to that of pure exploration? We'll do a more general analysis in the following section. Now, for intuition, suppose there's just  $K = 2$  arms, with Bernoulli reward distributions with means  $\mu^0 > \mu^1$ .

Let's let  $r^0$  be the random reward from the first arm and  $r^1$  be the random reward from the second. If  $r^0 > r^1$ , then we achieve zero regret. Otherwise, we achieve regret  $T(\mu^0 - \mu^1)$ . Thus, the expected regret is simply:

$$\begin{aligned} \mathbb{E}[\text{Regret}_T] &= \mathbb{P}(r^0 < r^1) \cdot T(\mu^0 - \mu^1) + c \\ &= (1 - \mu^0)\mu^1 \cdot T(\mu^0 - \mu^1) + c \end{aligned}$$

Which is still  $\Theta(T)$ , the same as pure exploration! Can we do better?

## 1.4 Explore-then-commit

We can improve the pure greedy algorithm as follows: let's reduce the variance of the reward estimates by pulling each arm  $> 1$  times before committing. This is called the **explore-then-commit** strategy.

### Definition 1.4.1: Explore-then-commit

```

Input:  $\leq T/K$ 
for  $k \leftarrow 0$  to  $K - 1$  do                                     ▷ Exploration phase
    for  $i \leftarrow 0$  to  $-1$  do
         $r_i^k \sim \nu^k$ 
    end for
     $\hat{\mu}^k \leftarrow \frac{1}{\sum_{i=0}^{-1}} r_i^k$ 
end for
 $\hat{k} \leftarrow \arg \max_k (\hat{\mu}^k)$ 
for  $t \leftarrow K$  to  $T - 1$  do                                     ▷ Exploitation phase
     $r_t \sim \nu^{\hat{k}}$ 
end for

```

(Note that the “pure greedy” strategy is just the special case where  $= 1$ .)

### 1.4.1 ETC regret analysis

Let's analyze the expected regret of this strategy by splitting it up into the exploration and exploitation phases.

**Exploration phase.** This phase takes  $K$  timesteps. Since at each step we incur at most 1 regret, the total regret is at most  $K$ .

**Exploitation phase.** This will take a bit more effort. We'll prove that for any total time  $T$ , we can choose  $K$  such that with arbitrarily high probability, the regret is sublinear. We know the regret from the exploitation phase is

$$T_{\text{exploit}}(\mu^* - \mu^{\hat{k}}) \quad \text{where} \quad T_{\text{exploit}} := T - K.$$

So we'd like to bound  $\mu^* - \mu^{\hat{k}} = o(1)$  (as a function of  $T$ ) in order to achieve sublinear regret. How can we do this?

Let's define  $\Delta^k = \hat{\mu}^k - \mu^k$  to denote how far the mean estimate for arm  $k$  is from the true mean. How can we bound this quantity? We'll use the following useful inequality for i.i.d. bounded random variables:

**Theorem 1.4.1: Hoeffding's inequality**

Let  $X_0, \dots, X_{n-1}$  be i.i.d. random variables with  $X_i \in [0, 1]$  almost surely for each  $i \in [n]$ . Then for any  $\delta > 0$ ,

$$\mathbb{P} \left( \left| \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \right| > \sqrt{\frac{\ln(2/\delta)}{2n}} \right) \leq \delta. \quad (1.2)$$

(The proof of this inequality is beyond the scope of this book.) We can apply this directly to the rewards for a given arm  $k$ , since the rewards from that arm are i.i.d.:

$$\mathbb{P} \left( |\Delta^k| > \sqrt{\frac{\ln(2/\delta)}{2}} \right) \leq \delta. \quad (1.3)$$

But note that we can't apply this to arm  $\hat{k}$  directly since  $\hat{k}$  is itself a random variable. Instead, we need to "uniform-ize" this bound across *all* the arms, i.e. bound the error across all the arms simultaneously, so that the resulting bound will apply *no matter what*  $\hat{k}$  "crystallizes" to.

The **union bound** provides a simple way to do this:

**Theorem 1.4.2: Union bound**

Consider a set of events  $A_0, \dots, A_{n-1}$ . Then

$$\mathbb{P}(\exists i \in [n]. A_i) \leq \sum_{i=0}^{n-1} \mathbb{P}(A_i).$$

In particular, if  $\mathbb{P}(A_i) \geq 1 - \delta$  for each  $i \in [n]$ , we have

$$\mathbb{P}(\forall i \in [n]. A_i) \geq 1 - n\delta.$$

**Exercise:** Prove the second statement above.

Applying the union bound across the arms for the l.h.s. event of ??, we have

$$\mathbb{P}\left(\forall k \in [K], |\Delta^k| \leq \sqrt{\frac{\ln(2/\delta)}{2}}\right) \geq 1 - K\delta$$

Then to apply this bound to  $\hat{k}$  in particular, we can apply the useful trick of “adding zero”:

$$\begin{aligned} \mu^{k^*} - \mu^{\hat{k}} &= \mu^{k^*} - \mu^{\hat{k}} + (\hat{\mu}^{k^*} - \hat{\mu}^{k^*}) + (\hat{\mu}^{\hat{k}} - \hat{\mu}^{\hat{k}}) \\ &= \Delta^{\hat{k}} - \Delta^{k^*} + \underbrace{(\hat{\mu}^{k^*} - \hat{\mu}^{\hat{k}})}_{\leq 0 \text{ by definition of } \hat{k}} \\ &\leq 2\sqrt{\frac{\ln(2K/\delta')}{2}} \text{ with probability at least } 1 - \delta' \end{aligned}$$

where we’ve set  $\delta' = K\delta$ . Putting this all together, we’ve shown that, with probability  $1 - \delta'$ ,

$$\text{Regret}_T \leq K + T_{\text{exploit}} \cdot \sqrt{\frac{2 \ln(2K/\delta')}{2}}.$$

Note that it suffices for to be on the order of  $\sqrt{T}$  to achieve sublinear regret. In particular, we can find the optimal by setting the derivative of the r.h.s. to zero:

$$\begin{aligned} 0 &= K - T_{\text{exploit}} \cdot \frac{1}{2} \sqrt{\frac{2 \ln(2K/\delta')}{3}} \\ &= \left( T_{\text{exploit}} \cdot \frac{\sqrt{\ln(2K/\delta')/2}}{K} \right)^{2/3} \end{aligned}$$

Plugging this into the expression for the regret, we have (still with probability  $1 - \delta'$ )

$$\begin{aligned} \text{Regret}_T &\leq 3T^{2/3} \sqrt[3]{K \ln(2K/\delta')/2} \\ &= \tilde{O}(T^{2/3} K^{1/3}). \end{aligned}$$

The ETC algorithm is rather “abrupt” in that it switches from exploration to exploitation after a fixed number of timesteps. In practice, it’s often better to use a more gradual transition, which brings us to the *epsilon-greedy* algorithm.

## 1.5 Epsilon-greedy

Instead of doing all of the exploration and then all of the exploitation separately – which additionally requires knowing the time horizon beforehand – we can instead interleave exploration and exploitation by, at each timestep, choosing a random action with some probability. We call this the **epsilon-greedy** algorithm.

### Definition 1.5.1: Epsilon-greedy

```

Input:  $\epsilon : \mathbb{N} \rightarrow [0, 1]$ 
 $S^k \leftarrow 0$  for each  $k \in [K]$ 
 $N^k \leftarrow 0$  for each  $k \in [K]$ 
for  $t \leftarrow 1$  to  $T$  do
  if  $\text{random}() < \epsilon(t)$  then
     $k \sim \text{Unif}([K])$ 
    ▷ Exploration
  else
     $k \leftarrow \arg \max_k \left( \frac{S^k}{N^k} \right)$ 
    ▷ Exploitation
  end if
   $r_t \sim \nu^k$ 
   $S^k \leftarrow S^k + r_t$ 
   $N^k \leftarrow N^k + 1$ 
end for

```

Note that we let  $\epsilon$  vary over time. In particular we might want to gradually *decrease*  $\epsilon$  as we learn more about the reward distributions over time.

It turns out that setting  $\epsilon_t = \sqrt[3]{K \ln(t)/t}$  also achieves a regret of  $\tilde{O}(t^{2/3} K^{1/3})$  (ignoring the logarithmic factors). (We will not prove this here.)

In ETC, we had to set based on the total number of timesteps  $T$ . But the epsilon-greedy algorithm actually handles the exploration *automatically*: the regret rate holds for *any*  $t$ , and doesn't depend on the final horizon  $T$ .

But the way these algorithms explore is rather naive: we've been exploring *uniformly* across all the arms. But what if we could be smarter about it, and explore *more* for arms that we're less certain about?

## 1.6 Upper Confidence Bound (UCB)

To quantify how *certain* we are about the mean of each arm, we'll compute *confidence intervals* for our estimators, and then choose the arm with the highest *upper confidence bound*. This operates on the principle of **the benefit of the doubt (i.e. optimism in the face of uncertainty)**: we'll choose the arm that we're most optimistic about.

In particular, for each arm  $k$  at time  $t$ , we'd like to compute some upper confidence bound  $M_t^k$  such that  $\hat{\mu}_t^k \leq M_t^k$  with high probability, and then choose  $a_t := \arg \max_{k \in [K]} M_t^k$ . But how

should we compute  $M_t^k$ ?

In ??, we were able to compute this bound using Hoeffding's inequality, which assumes that the number of samples is *fixed*. This was the case in ETC (where we pull each arm  $t$  times), but in UCB, the number of times we pull each arm depends on the agent's actions, which in turn depend on the random rewards and are therefore stochastic. So we *can't* use Hoeffding's inequality directly.

Instead, we'll apply the same trick we used in the ETC analysis: we'll use the **union bound** to compute a *looser* bound that holds *uniformly* across all timesteps and arms. Let's introduce some notation to discuss this.

Let  $N_t^k$  denote the (random) number of times arm  $k$  has been pulled within the first  $t$  timesteps, and  $\hat{\mu}_t^k$  denote the sample average of those pulls. That is,

$$N_t^k := \sum_{\tau=0}^{t-1} \mathbf{1}\{a_\tau = k\}$$

$$\hat{\mu}_t^k := \frac{1}{N_t^k} \sum_{\tau=0}^{t-1} \mathbf{1}\{a_\tau = k\} r_\tau.$$

To achieve the “fixed sample size” assumption, we'll need to shift our index from *time* to *number of samples from each arm*. In particular, we'll define  $\tilde{r}_n^k$  to be the  $n$ th sample from arm  $k$ , and  $\tilde{\mu}_n^k$  to be the sample average of the first  $n$  samples from arm  $k$ . Then, for a fixed  $n$ , this satisfies the “fixed sample size” assumption, and we can apply Hoeffding's inequality to get a bound on  $\tilde{\mu}_n^k$ .

So how can we extend our bound on  $\tilde{\mu}_n^k$  to  $\hat{\mu}_t^k$ ? Well, we know  $N_t^k \leq t$  (where equality would be the case if and only if we had pulled arm  $k$  every time). So we can apply the same trick as last time, where we uniform-ize across all possible values of  $N_t^k$ :

$$\mathbb{P} \left( \forall n \leq t, |\tilde{\mu}_n^k - \mu^k| \leq \sqrt{\frac{\ln(2/\delta)}{2n}} \right) \geq 1 - t\delta.$$

In particular, since  $N_t^k \leq t$ , and  $\tilde{\mu}_{N_t^k}^k = \hat{\mu}_t^k$  by definition, we have

$$\mathbb{P} \left( |\hat{\mu}_t^k - \mu^k| \leq \sqrt{\frac{\ln(2t/\delta')}{2N_t^k}} \right) \geq 1 - \delta' \text{ where } \delta' := t\delta.$$

This bound would then suffice for applying the UCB algorithm! That is, the upper confidence bound for arm  $k$  would be

$$M_t^k := \hat{\mu}_t^k + \sqrt{\frac{\ln(2t/\delta')}{2N_t^k}},$$

where we can choose  $\delta'$  depending on how tight we want the interval to be. A smaller  $\delta'$  would give us a larger yet higher-confidence interval, and vice versa. We can now use this to define the UCB algorithm.

**Definition 1.6.1: Upper Confidence Bound (UCB)**

**Input:**  $\delta' \in (0, 1)$   
**for**  $t \leftarrow 0$  to  $T - 1$  **do**  
 $k \leftarrow \arg \max_{k' \in [K]} \frac{S^{k'}}{N^{k'}} + \sqrt{\frac{\ln(2t/\delta')}{2N^{k'}}}$   
 $r_t \sim \nu^k$   
 $S^k \leftarrow S^k + r_t$   
 $N^k \leftarrow N^k + 1$   
**end for**

**Exercise:** As written, this ignores the issue that we divide by  $N^k = 0$  for all arms at the beginning. How should we resolve this issue?

Intuitively, UCB prioritizes arms where:

1.  $\hat{\mu}_t^k$  is large, i.e. the arm has a high sample average, and we'd choose it for *exploitation*, and
2.  $\sqrt{\frac{\ln(2t/\delta')}{2N_t^k}}$  is large, i.e. we're still uncertain about the arm, and we'd choose it for *exploration*.

As desired, this explores in a smarter, *adaptive* way compared to the previous algorithms. Does it achieve lower regret?

### 1.6.1 UCB regret analysis

First we'll bound the regret incurred at each timestep. Then we'll bound the *total* regret across timesteps.

For the sake of analysis, we'll use a slightly looser bound that applies across the whole time horizon and across all arms. We'll omit the derivation since it's very similar to the above (walk through it yourself for practice).

$$\mathbb{P}(\forall k \leq K, t < T. |\hat{\mu}_t^k - \mu^k| \leq B_t^k) \geq 1 - \delta''$$

$$\text{where } B_t^k := \sqrt{\frac{\ln(2TK/\delta'')}{2N_t^k}}.$$

Intuitively,  $B_t^k$  denotes the *width* of the CI for arm  $k$  at time  $t$ . Then, assuming the above uniform bound holds (which occurs with probability  $1 - \delta''$ ), we can bound the regret at each timestep as follows:

$$\begin{aligned} \mu^* - \mu^{a_t} &\leq \hat{\mu}_t^{k^*} + B_t^{k^*} - \mu^{a_t} && \text{applying UCB to arm } k^* \\ &\leq \hat{\mu}_t^{a_t} + B_t^{a_t} - \mu^{a_t} && \text{since UCB chooses } a_t = \arg \max_{k \in [K]} \hat{\mu}_t^k + B_t^k \end{aligned}$$

$$\leq 2B_t^{a_t} \quad \text{since } \hat{\mu}_t^{a_t} - \mu^{a_t} \leq B_t^{a_t} \text{ by definition of } B_t^{a_t}$$

Summing this across timesteps gives

$$\begin{aligned} \text{Regret}_T &\leq \sum_{t=0}^{T-1} 2B_t^{a_t} \\ &= \sqrt{2 \ln(2TK/\delta'')} \sum_{t=0}^{T-1} (N_t^{a_t})^{-1/2} \\ \sum_{t=0}^{T-1} (N_t^{a_t})^{-1/2} &= \sum_{t=0}^{T-1} \sum_{k=1}^K \mathbf{1}\{a_t = k\} (N_t^k)^{-1/2} \\ &= \sum_{k=1}^K \sum_{n=1}^{N_T^k} n^{-1/2} \\ &\leq K \sum_{n=1}^T n^{-1/2} \\ \sum_{n=1}^T n^{-1/2} &\leq 1 + \int_1^T x^{-1/2} dx \\ &= 1 + (2\sqrt{x})_1^T \\ &= 2\sqrt{T} - 1 \\ &\leq 2\sqrt{T} \end{aligned}$$

Putting everything together gives

$$\begin{aligned} \text{Regret}_T &\leq 2K \sqrt{2T \ln(2TK/\delta'')} && \text{with probability } 1 - \delta'' \\ &= \tilde{O}(K\sqrt{T}) \end{aligned}$$

In fact, we can do a more sophisticated analysis to trim off a factor of  $\sqrt{K}$  and show  $\text{Regret}_T = \tilde{O}(\sqrt{TK})$ .

### 1.6.2 Lower bound on regret (intuition)

Is it possible to do better than  $\Omega(\sqrt{T})$  in general? In fact, no! We can show that any algorithm must incur  $\Omega(\sqrt{T})$  regret in the worst case. We won't rigorously prove this here, but the intuition is as follows.

The Central Limit Theorem tells us that with  $T$  i.i.d. samples from some distribution, we can only learn the mean of the distribution to within  $\Omega(1/\sqrt{T})$  (the standard deviation). Then, since

we get  $T$  samples spread out across the arms, we can only learn each arm's mean to an even looser degree.

That is, if two arms have means that are within about  $1/\sqrt{T}$ , we won't be able to confidently tell them apart, and will sample them about equally. But then we'll incur regret

$$\Omega((T/2) \cdot (1/\sqrt{T})) = \Omega(\sqrt{T}).$$

## 1.7 Thompson sampling and Bayesian bandits

So far, we've treated the parameters  $\mu^0, \dots, \mu^{K-1}$  of the reward distributions as *fixed*. Instead, we can take a **Bayesian** approach where we treat them as random variables from some **prior distribution**. Then, upon pulling an arm and observing a reward, we can simply *condition* on this observation to exactly describe the **posterior distribution** over the parameters. This fully describes the information we gain about the parameters from observing the reward.

From this Bayesian perspective, the **Thompson sampling** algorithm follows naturally: just sample from the distribution of the optimal arm, given the observations!

### Definition 1.7.1: Thompson sampling

**Input:** the prior distribution  $\pi \in \Delta([0, 1]^K)$   
**for**  $t \leftarrow 0$  to  $T - 1$  **do**  
      $\sim \pi(\cdot \mid a_0, r_0, \dots, a_{t-1}, r_{t-1})$   
      $a_t \leftarrow \arg \max_{k \in [K]} \mu^k$   
      $r_t \sim \nu^{a_t}$  ▷ Observe reward  
**end for**

In other words, we sample each arm proportionally to how likely we think it is to be optimal, given the observations so far. This strikes a good exploration-exploitation tradeoff: we explore more for arms that we're less certain about, and exploit more for arms that we're more certain about. Thompson sampling is a simple yet powerful algorithm that achieves state-of-the-art performance in many settings.

### Example 1.7.1: Bayesian Bernoulli bandit

We've often been working in the Bernoulli bandit setting, where arm  $k$  yields a reward of 1 with probability  $\mu^k$  and no reward otherwise. The vector of success probabilities  $= (\mu^1, \dots, \mu^K)$  thus describes the entire MAB.

Under the Bayesian perspective, we think of  $\mu$  as a *random* vector drawn from some prior distribution  $\pi()$ . For example, we might have  $\pi$  be the Uniform distribution over the unit hypercube  $[0, 1]^K$ , that is,

$$\pi() = \begin{cases} 1 & \text{if } \mu \in [0, 1]^K \\ 0 & \text{otherwise} \end{cases}$$



Then, upon viewing some reward, we can exactly calculate the **posterior** distribution of using Bayes's rule (i.e. the definition of conditional probability):

$$\begin{aligned}\mathbb{P}(a_0, r_0) &\propto \mathbb{P}(r_0 | a_0) \mathbb{P}(a_0) \\ &\propto (\mu^{a_0})^{r_0} (1 - \mu^{a_0})^{1-r_0}.\end{aligned}$$

This is the PDF of the  $\text{Beta}(1 + r_0, 1 + (1 - r_0))$  distribution, which is a conjugate prior for the Bernoulli distribution. That is, if we start with a Beta prior on  $\mu^k$  (note that  $\text{Unif}([0, 1]) = \text{Beta}(1, 1)$ ), then the posterior, after conditioning on samples from  $\text{Bern}(\mu^k)$ , will also be Beta. This is a very convenient property, since it means we can simply update the parameters of the Beta distribution upon observing a reward, rather than having to recompute the entire posterior distribution from scratch.

It turns out that asymptotically, Thompson sampling is optimal in the following sense. Lai and Robbins [**lai asymptotically 1985**] prove an *instance-dependent* lower bound that says for *any* bandit algorithm,

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[N_T^k]}{\ln(T)} \geq \frac{1}{\text{KL}(\mu^k \parallel \mu^*)}$$

where

$$\text{KL}(\mu^k \parallel \mu^*) = \mu^k \ln \frac{\mu^k}{\mu^*} + (1 - \mu^k) \ln \frac{1 - \mu^k}{1 - \mu^*}$$

measures the **Kullback-Leibler divergence** from the Bernoulli distribution with mean  $\mu^k$  to the Bernoulli distribution with mean  $\mu^*$ . It turns out that Thompson sampling achieves this lower bound with equality! That is, not only is the error *rate* optimal, but the *constant factor* is optimal as well.

## **Chapter 2**

# **Markov Decision Processes**

# Contents

## 2.1 Introduction

The field of RL studies how an agent can learn to make sequential decisions in an interactive environment. This is a very general problem! How can we *formalize* this task in a way that is both *sufficiently general* yet also tractable enough for *fruitful analysis*?

Let's consider some examples of sequential decision problems to identify the key common properties we'd like to capture:

- **Board games** like chess or Go, where the player takes turns with the opponent to make moves on the board.
- **Video games** like Super Mario Bros or Breakout, where the player controls a character to reach the goal.
- **Robotic control**, where the robot can move and interact with the real-world environment to complete some task.

All of these fit into the RL framework. Furthermore, these are environments where the **state transitions**, the “rules” of the environment, only depend on the *most recent* state and action. This is called the **Markov property**.

### Definition 2.1.1: Markov property

An interactive environment satisfies the **Markov property** if the probability of transitioning to a new state only depends on the current state and action:

$$\mathbb{P}(s_{h+1} \mid s_0, a_0, \dots, s_h, a_h) = P(s_{h+1} \mid s_h, a_h)$$

where  $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$  describes the state transitions. (We'll elaborate on this notation later in the chapter.)

We'll see that this simple assumption leads to a rich set of problems and algorithms. Environments with the Markov property are called **Markov decision processes** (MDPs) and will be the focus of this chapter.

**Exercise:** What information might be encoded in the state for each of the above examples? What might the valid set of actions be? Describe the state transitions heuristically and verify that they satisfy the Markov property.

MDPs are usually classified as **finite-horizon**, where the interactions end after some finite number of time steps, or **infinite-horizon**, where the interactions can continue indefinitely. We'll begin with the finite-horizon case and discuss the infinite-horizon case in the second half of the chapter.

In each setting, we'll describe how to evaluate different **policies** (strategies for choosing actions) and how to compute (or approximate) the **optimal policy** for a given MDP. We'll introduce the **Bellman consistency condition**, which allows us to analyze the whole series of interactions in terms of individual timesteps.

## 2.2 Finite horizon MDPs

### Definition 2.2.1: Finite-horizon Markov decision process

The components of a finite-horizon Markov decision process are:

1. The **state** that the agent interacts with. We use  $\mathcal{S}$  to denote the set of possible states, called the **state space**.
2. The **actions** that the agent can take. We use  $\mathcal{A}$  to denote the set of possible actions, called the **action space**.
3. Some **initial state distribution**  $\mu \in \Delta(\mathcal{S})$ .
4. The **state transitions** (a.k.a. **dynamics**)  $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$  that describe what state the agent transitions to after taking an action.
5. The **reward** signal. In this course we'll take it to be a deterministic function on state-action pairs,  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ , but in general many results will extend to a *stochastic* reward signal.
6. A time horizon  $H \in \mathbb{N}$  that specifies the number of interactions in an **episode**.

Combined together, these objects specify a finite-horizon Markov decision process:

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

### Example 2.2.1: Tidying MDP

Let's consider an extremely simple decision problem throughout this chapter: the task of keeping your room tidy!

Your room has the possible states  $\mathcal{S} = \{\text{orderly}, \text{messy}\}$ . You can take either of the actions  $\mathcal{A} = \{\text{tidy}, \text{ignore}\}$ . The room starts off orderly.

The state transitions are as follows: if you tidy the room, it becomes (or remains) orderly; if you ignore the room, it might become messy.

The rewards are as follows: You get penalized for tidying an orderly room (a waste of time) or ignoring a messy room, but you get rewarded for ignoring an orderly room (since you can enjoy). Tidying a messy room is a chore that gives no reward.

These are summarized in the following table:

$s$	$a$	$P(\text{orderly} \mid s, a)$	$P(\text{messy} \mid s, a)$	$r(s, a)$
orderly	tidy	1	0	-1
orderly	ignore	0.7	0.3	1
messy	tidy	1	0	0
messy	ignore	0	1	-1

Consider a time horizon of  $H = 7$  days (one interaction per day). Let  $t = 0$  correspond to Monday and  $t = 6$  correspond to Sunday.

## 2.2.1 Policies

### Definition 2.2.2: Policies

A **policy**  $\pi$  describes the agent's strategy: which actions it takes in a given situation. A key goal of RL is to find the **optimal policy** that maximizes the total reward on average.

There are three axes along which policies can vary: their outputs, inputs, and time-dependence. We'll discuss each of these in turn.

1. **Deterministic or stochastic.** A deterministic policy outputs actions while a stochastic policy outputs *distributions* over actions.
2. **State-dependent or history-dependent.** A state-dependent (a.k.a. "Markovian") policy only depends on the current state, while a history-dependent policy depends on the sequence of past states, actions, and rewards. We'll only consider state-dependent policies in this course.
3. **Stationary or time-dependent.** A stationary policy remains the same function at all time steps, while a time-dependent policy  $\pi = \{\pi_0, \dots, \pi_{H-1}\}$  specifies a different function  $\pi_h$  at each time step  $h$ .

A fascinating result is that every finite-horizon MDP has an optimal deterministic time-dependent policy! Intuitively, the Markov property implies that the current state contains all the information we need to make the optimal decision. We'll prove this result constructively later in the chapter.

### Example 2.2.2: Tidying policies

Here are some possible policies for the tidying MDP (??):

- Always tidy:  $\pi(s) = \text{tidy}$ .
- Only tidy on weekends:  $\pi_h(s) = \text{tidy}$  if  $h \in \{5, 6\}$  and  $\pi_h(s) = \text{ignore}$  otherwise.
- Only tidy if the room is messy:  $\pi_h(\text{messy}) = \text{tidy}$  and  $\pi_h(\text{orderly}) = \text{ignore}$  for all  $h$ .

## 2.2.2 Trajectories

### Definition 2.2.3: Trajectories

A sequence of states, actions, and rewards is called a **trajectory**:

$$\tau = (s_0, a_0, r_0, \dots, s_{H-1}, a_{H-1}, r_{H-1})$$

where  $r_h = r(s_h, a_h)$ . (Note that sources differ as to whether to include the reward at the final time step. This is a minor detail.)

Once we've chosen a policy, we can sample trajectories by repeatedly choosing actions according to the policy, transitioning according to the state transitions, and observing the rewards. That is, a policy induces a distribution  $\rho^\pi$  over trajectories. (We assume that  $\mu$  and  $P$  are clear from context.)

### Example 2.2.3: Trajectories in the tidying environment

Here is a possible trajectory for the tidying example:

$t$	0	1	2	3	4	5	6
$s$	orderly	orderly	orderly	messy	messy	orderly	orderly
$a$	tidy	ignore	ignore	ignore	tidy	ignore	ignore
$r$	-1	1	1	-1	0	1	1

Could any of the policies in ?? have generated this trajectory?

Note that for a state-dependent policy, using the Markov property (??), we can specify this probability distribution in an **autoregressive** way (i.e. one timestep at a time):

### Definition 2.2.4: Autoregressive trajectory distribution

$$\rho^\pi(\tau) := \mu(s_0)\pi_0(a_0 | s_0)P(s_1 | s_0, a_0) \cdots P(s_{H-1} | s_{H-2}, a_{H-2})\pi_{H-1}(a_{H-1} | s_{H-1})$$

**Exercise:** How would you modify this to include stochastic rewards?

For a deterministic policy  $\pi$ , we have that  $\pi_h(a | s) = \mathbb{I}[a = \pi_h(s)]$ ; that is, the probability of taking an action is 1 if it's the unique action prescribed by the policy for that state and 0 otherwise. In this case, the only randomness in sampling trajectories comes from the initial state distribution  $\mu$  and the state transitions  $P$ .

## 2.2.3 Value functions

The main goal of RL is to find a policy that maximizes the average total reward  $r_0 + \cdots + r_{H-1}$ . (Note that this is a random variable that depends on the policy.) Let's introduce some notation for analyzing this quantity.

A policy's **value function** is its expected total reward *starting in a given state at a given*

time:

**Definition 2.2.5: Value function**

$$V_h^\pi(s) := \mathbb{E}_{\tau \sim \rho^\pi} [r_h + \dots + r_{H-1} \mid s_h = s]$$

Similarly, we can define the **action-value function** (aka the **Q-function**) as the expected total reward when starting in a given state and taking a given action:

**Definition 2.2.6: Action-value function**

$$Q_h^\pi(s, a) := \mathbb{E}_{\tau \sim \rho^\pi} [r_h + \dots + r_{H-1} \mid s_h = s, a_h = a]$$

Note that the value function is just the average action-value over actions drawn from the policy:

$$V_h^\pi(s) = \mathbb{E}_{a \sim \pi_h(s)} [Q_h^\pi(s, a)]$$

and the action-value can be expressed in terms of the value of the following state:

$$Q_h^\pi(s, a) = r(s, a) + \mathbb{E}_{s' \sim P(s, a)} [V_{h+1}^\pi(s')]$$

### The one-step (Bellman) consistency equation

Note that by simply considering the cumulative reward as the sum of the *current* reward and the *future* cumulative reward, we can describe the value function recursively (in terms of itself). This is named the **Bellman consistency equation** after **Richard Bellman** (1920–1984), who is credited with introducing dynamic programming in 1953.

**Definition 2.2.7: Bellman consistency equation for the value function**

$$V_h^\pi(s) = \mathbb{E}_{\substack{a \sim \pi_h(s) \\ s' \sim P(s, a)}} [r(s, a) + V_{h+1}^\pi(s')] \quad (2.1)$$

**Exercise:** Verify that this equation holds by expanding  $V_h^\pi(s)$  and  $V_{h+1}^\pi(s')$ .

One can analogously derive the Bellman consistency equation for the action-value function:

**Definition 2.2.8: Bellman consistency equation for action-values**

$$Q_h^\pi(s, a) = r(s, a) + \mathbb{E}_{\substack{s' \sim P(s, a) \\ a' \sim \pi_{h+1}(s')}} [Q_{h+1}^\pi(s', a')]$$



**Remark 2.2.1: The Bellman consistency equation for deterministic policies**

Note that for deterministic policies, the Bellman consistency equation simplifies to

$$\begin{aligned} V_h^\pi(s) &= r(s, \pi_h(s)) + \mathbb{E}_{s' \sim P(s, \pi_h(s))} [V_{h+1}^\pi(s')] \\ Q_h^\pi(s, a) &= r(s, a) + \mathbb{E}_{s' \sim P(s, a)} [Q_{h+1}^\pi(s', \pi_{h+1}(s'))] \end{aligned}$$

**The Bellman operator**

Fix a policy  $\pi$ . Consider the higher-order operator that takes in a “value function”  $v : \mathcal{S} \rightarrow \mathbb{R}$  and returns the r.h.s. of the Bellman equation for that “value function”:

**Definition 2.2.9: Bellman operator**

$$[\mathcal{J}^\pi(v)](s) := \mathbb{E}_{\substack{a \sim \pi(s) \\ s' \sim P(s, a)}} [r(s, a) + v(s')].$$

We’ll call  $\mathcal{J}^\pi : (\mathcal{S} \rightarrow \mathbb{R}) \rightarrow (\mathcal{S} \rightarrow \mathbb{R})$  the **Bellman operator** of  $\pi$ . Note that it’s defined on any “value function” mapping states to real numbers;  $v$  doesn’t have to be a well-defined value function for some policy (hence the lowercase notation). The Bellman operator also gives us a concise way to express the Bellman consistency equation (??):

$$V_h^\pi = \mathcal{J}^\pi(V_{h+1}^\pi)$$

Intuitively, the output of the Bellman operator, a new “value function”, evaluates states as follows: from a given state, take one action according to  $\pi$ , observe the reward, and then evaluate the next state using the input “value function”.

When we discuss infinite-horizon MDPs, the Bellman operator will turn out to be more than just a notational convenience: We’ll use it to construct algorithms for computing the optimal policy.

**2.2.4 Policy evaluation**

How can we actually compute the value function of a given policy? This is the task of **policy evaluation**.

**Dynamic programming**

The Bellman consistency equation (??) gives us a convenient algorithm for evaluating stationary policies: it expresses the value function at timestep  $h$  as a function of the value function at timestep  $h + 1$ . This means we can start at the end of the time horizon, where the value is known, and work backwards in time, using the Bellman consistency equation to compute the value function at each time step.

**Definition 2.2.10: Dynamic programming for policy evaluation**

```

 $V_h(s) \leftarrow 0$  for all  $t \in \{0, \dots, H\}, s \in \mathcal{S}$ 
for  $t = H - 1, \dots, 0$  do
  for  $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$  do
     $V_h(s) \leftarrow V_h(s) + \pi_h(a \mid s)P(s' \mid s, a)[r(s, a) + V_{h+1}(s')]$ 
  end for
end for

```

This clearly runs in time  $O(H \cdot |\mathcal{S}|^2 \cdot |\mathcal{A}|)$  by counting the loops.

**Exercise:** Do you see where we compute  $Q_h^\pi$  along the way? Make this step explicit.

**Example 2.2.4: Tidying policy evaluation**

Let's evaluate the policy from ?? that tidies if and only if the room is messy. We'll use the Bellman consistency equation to compute the value function at each time step.

$$\begin{aligned}
 V_{H-1}^\pi(\text{orderly}) &= r(\text{orderly}, \text{ignore}) \\
 &= 1 \\
 V_{H-1}^\pi(\text{messy}) &= r(\text{messy}, \text{tidy}) \\
 &= 0 \\
 V_{H-2}^\pi(\text{orderly}) &= r(\text{orderly}, \text{ignore}) + \mathbb{E}_{s' \sim P(\text{orderly}, \text{ignore})} [V_{H-1}^\pi(s')] \\
 &= 1 + 0.7 \cdot V_{H-1}^\pi(\text{orderly}) + 0.3 \cdot V_{H-1}^\pi(\text{messy}) \\
 &= 1 + 0.7 \cdot 1 + 0.3 \cdot 0 \\
 &= 1.7 \\
 V_{H-2}^\pi(\text{messy}) &= r(\text{messy}, \text{tidy}) + \mathbb{E}_{s' \sim P(\text{messy}, \text{tidy})} [V_{H-1}^\pi(s')] \\
 &= 0 + 1 \cdot V_{H-1}^\pi(\text{orderly}) + 0 \cdot V_{H-1}^\pi(\text{messy}) \\
 &= 1 \\
 V_{H-3}^\pi(\text{orderly}) &= r(\text{orderly}, \text{ignore}) + \mathbb{E}_{s' \sim P(\text{orderly}, \text{ignore})} [V_{H-2}^\pi(s')] \\
 &= 1 + 0.7 \cdot V_{H-2}^\pi(\text{orderly}) + 0.3 \cdot V_{H-2}^\pi(\text{messy}) \\
 &= 1 + 0.7 \cdot 1.7 + 0.3 \cdot 1 \\
 &= 2.49 \\
 V_{H-3}^\pi(\text{messy}) &= r(\text{messy}, \text{tidy}) + \mathbb{E}_{s' \sim P(\text{messy}, \text{tidy})} [V_{H-2}^\pi(s')] \\
 &= 0 + 1 \cdot V_{H-2}^\pi(\text{orderly}) + 0 \cdot V_{H-2}^\pi(\text{messy}) \\
 &= 1.7
 \end{aligned}$$

etc. You may wish to repeat this computation for the other policies to get a better sense of this algorithm.

### 2.2.5 Optimal policies

We've just seen how to *evaluate* a given policy. But how can we find the **optimal policy** for a given environment?

#### Definition 2.2.11: Optimal policies

We call a policy optimal, and denote it by  $\pi^*$ , if it does at least as well as *any* other policy  $\pi$  (including stochastic and history-dependent ones) in all situations:

$$\begin{aligned} V_h^{\pi^*}(s) &= \mathbb{E}_{\tau \sim \rho^{\pi^*}}[r_h + \dots + r_{H-1} \mid s_h = s] \\ &\geq \mathbb{E}_{\tau \sim \rho^\pi}[r_h + \dots + r_{H-1} \mid \tau_h] \quad \forall \pi, \tau_h, h \in [H] \end{aligned} \quad (2.2)$$

where we condition on the trajectory up to time  $h$ , denoted  $\tau_h = (s_0, a_0, r_0, \dots, s_h)$ , where  $s_h = s$ .

Convince yourself that all optimal policies must have the same value function. We call this the **optimal value function** and denote it by  $V_h^*(s)$ . The same goes for the action-value function  $Q_h^*(s, a)$ .

It is a stunning fact that **every finite-horizon MDP has an optimal policy that is time-dependent and deterministic**. In particular, we can construct such a policy by acting *greedily* with respect to the optimal action-value function:

$$\pi_h^*(s) = \arg \max_a Q_h^*(s, a).$$

#### Theorem 2.2.1: It is optimal to be greedy w.r.t. the optimal value function

Let  $V^*$  and  $Q^*$  denote the optimal value and action-value functions. Consider the greedy policy

$$\hat{\pi}_h(s) := \arg \max_a Q_h^*(s, a).$$

We aim to show that  $\hat{\pi}$  is optimal; that is,  $V^{\hat{\pi}} = V^*$ .

Fix an arbitrary state  $s \in \mathcal{S}$  and time  $h \in [H]$ .

Firstly, by the definition of  $V^*$ , we already know  $V_h^*(s) \geq V_h^{\hat{\pi}}(s)$ . So for equality to hold we just need to show that  $V_h^*(s) \leq V_h^{\hat{\pi}}(s)$ . We'll first show that the Bellman operator  $\mathcal{J}^{\hat{\pi}}$  never decreases  $V_h^*$ . Then we'll apply this result recursively to show that  $V^* = V^{\hat{\pi}}$ .

**Lemma:**  $\mathcal{J}^{\hat{\pi}}$  never decreases  $V_h^*$  (elementwise):

$$[\mathcal{J}^{\hat{\pi}}(V_{h+1}^*)](s) \geq V_h^*(s).$$

**Proof:**

$$\begin{aligned}
V_h^*(s) &= \max_{\pi \in \Pi} V_h^\pi(s) \\
&= \max_{\pi \in \Pi} \mathbb{E}_{a \sim \pi(\dots)} \left[ r(s, a) + \mathbb{E}_{s' \sim P(s, a)} V_{h+1}^\pi(s') \right] && \text{Bellman consistency} \\
&\leq \max_{\pi \in \Pi} \mathbb{E}_{a \sim \pi(\dots)} \left[ r(s, a) + \mathbb{E}_{s' \sim P(s, a)} V_{h+1}^*(s') \right] && \text{definition of } V^* \\
&= \max_a \left[ r(s, a) + \mathbb{E}_{s' \sim P(s, a)} V_{h+1}^*(s') \right] && \text{only depends on } \pi \text{ via } a \\
&= [\mathcal{J}^{\hat{\pi}}(V_{h+1}^*)](s).
\end{aligned}$$

Note that the chosen action  $a \sim \pi(\dots)$  above might depend on the past history; this isn't shown in the notation and doesn't affect our result (make sure you see why).

We can now apply this result recursively to get

$$V_t^*(s) \leq V_t^{\hat{\pi}}(s)$$

as follows. (Note that even though  $\hat{\pi}$  is deterministic, we'll use the  $a \sim \hat{\pi}(s)$  notation to make it explicit that we're sampling a trajectory from it.)

$$\begin{aligned}
V_t^*(s) &\leq [\mathcal{J}^{\hat{\pi}}(V_{h+1}^*)](s) \\
&= \mathbb{E}_{a \sim \hat{\pi}(s)} \left[ r(s, a) + \mathbb{E}_{s' \sim P(s, a)} [V_{h+1}^*(s')] \right] && \text{definition of } \mathcal{J}^{\hat{\pi}} \\
&\leq \mathbb{E}_{a \sim \hat{\pi}(s)} \left[ r(s, a) + \mathbb{E}_{s' \sim P(s, a)} [[\mathcal{J}^{\hat{\pi}}(V_{t+2}^*)](s')] \right] && \text{above lemma} \\
&= \mathbb{E}_{a \sim \hat{\pi}(s)} \left[ r(s, a) + \mathbb{E}_{s' \sim P(s, a)} \left[ \mathbb{E}_{a' \sim \hat{\pi}} r(s', a') + \mathbb{E}_{s'' \sim P(s', a')} V_{t+2}^*(s'') \right] \right] && \text{definition of } \mathcal{J}^{\hat{\pi}} \\
&\leq \dots && \text{apply at all timesteps} \\
&= \mathbb{E}_{\tau \sim \rho^{\hat{\pi}}} [G_t \mid s_h = s] && \text{rewrite expectation} \\
&= V_t^{\hat{\pi}}(s) && \text{definition}
\end{aligned}$$

And so we have  $V^* = V^{\hat{\pi}}$ , making  $\hat{\pi}$  optimal.

## Dynamic programming

Now that we've shown this particular greedy policy is optimal, all we need to do is compute the optimal value function and optimal policy. We can do this by working backwards in time using **dynamic programming** (DP).

**Definition 2.2.12: DP for optimal policy**

We can solve for the optimal policy in an finite-horizon MDP using **dynamic programming**.

- *Base case.* At the end of the episode (time step  $H - 1$ ), we can't take any more actions, so the  $Q$ -function is simply the reward that we obtain:

$$Q_{H-1}^*(s, a) = r(s, a)$$

so the best thing to do is just act greedily and get as much reward as we can!

$$\pi_{H-1}^*(s) = \arg \max_a Q_{H-1}^*(s, a)$$

Then  $V_{H-1}^*(s)$ , the optimal value of state  $s$  at the end of the trajectory, is simply whatever action gives the most reward.

$$V_{H-1}^* = \max_a Q_{H-1}^*(s, a)$$

- *Recursion.* Then, we can work backwards in time, starting from the end, using our consistency equations! i.e. for each  $t = H - 2, \dots, 0$ , we set

$$Q_t^*(s, a) = r(s, a) + \mathbb{E}_{s' \sim P(s, a)} [V_{t+1}^*(s')]$$

$$\pi_t^*(s) = \arg \max_a Q_t^*(s, a)$$

$$V_t^*(s) = \max_a Q_t^*(s, a)$$

At each of the  $H$  timesteps, we must compute  $Q^*$  for each of the  $|\mathcal{S}||\mathcal{A}|$  state-action pairs. Each computation takes  $|\mathcal{S}|$  operations to evaluate the average value over  $s'$ . This gives a total computation time of  $O(H|\mathcal{S}|^2|\mathcal{A}|)$ .

Note that this algorithm is identical to the policy evaluation algorithm ??, but instead of *averaging* over the actions chosen by a policy, we instead simply take a *maximum* over the action-values. We'll see this relationship between **policy evaluation** and **optimal policy computation** show up again in the infinite-horizon setting.

**Example 2.2.5: Optimal policy for the tidying MDP**

Left as an exercise.

## 2.3 Infinite horizon MDPs

What happens if a trajectory is allowed to continue forever (i.e.  $H = \infty$ )? This is the setting of **infinite horizon** MDPs.

In this chapter, we'll describe the necessary adjustments from the finite-horizon case to make the problem tractable. We'll show that the Bellman operator (??) in the discounted reward setting is a **contraction mapping** for any policy. We'll discuss how to evaluate policies (i.e. compute their corresponding value functions). Finally, we'll present and analyze two iterative algorithms, based on the Bellman operator, for computing the optimal policy: **value iteration** and **policy iteration**.

### 2.3.1 Differences from finite-horizon

#### Discounted rewards

First of all, note that maximizing the cumulative reward  $r_h + r_{h+1} + r_{h+2} + \dots$  is no longer a good idea since it might blow up to infinity. Instead of a time horizon  $H$ , we now need a **discount factor**  $\gamma \in [0, 1)$  such that rewards become less valuable the further into the future they are:

$$r_h + \gamma r_{h+1} + \gamma^2 r_{h+2} + \dots = \sum_{k=0}^{\infty} \gamma^k r_{h+k}.$$

We can think of  $\gamma$  as measuring how much we care about the future: if it's close to 0, we only care about the near-term rewards; it's close to 1, we put more weight into future rewards.

You can also analyze  $\gamma$  as the probability of *continuing* the trajectory at each time step. (This is equivalent to  $H$  being distributed by a First Success distribution with success probability  $\gamma$ .) This accords with the above interpretation: if  $\gamma$  is close to 0, the trajectory will likely be very short, while if  $\gamma$  is close to 1, the trajectory will likely continue for a long time.

**Exercise:** Assuming that  $r_h \in [0, 1]$  for all  $h \in \mathbb{N}$ , what is the maximum **discounted** cumulative reward? You may find it useful to review geometric series.

The other components of the MDP remain the same:

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

#### Stationary policies

The time-dependent policies from the finite-horizon case become difficult to handle in the infinite-horizon case. In particular, many of the DP approaches we saw required us to start at the end of the trajectory, which is no longer possible. We'll shift to **stationary** policies  $\pi : \mathcal{S} \rightarrow \mathcal{A}$  (deterministic) or  $\Delta(\mathcal{A})$  (stochastic).

**Exercise:** Which of the policies in ?? are stationary?

### Value functions and Bellman consistency

We also consider stationary value functions  $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$  and  $Q^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ . We need to insert a factor of  $\gamma$  into the Bellman consistency equation (??) to account for the discounting:

$$\begin{aligned}
 V^\pi(s) &= \mathbb{E}_{\tau \sim \rho^\pi} [r_h + \gamma r_{h+1} + \gamma^2 r_{h+2} + \dots \mid s_h = s] && \text{for any } h \in \mathbb{N} \\
 &= \mathbb{E}_{\substack{a \sim \pi(s) \\ s' \sim P(s,a)}} [r(s, a) + \gamma V^\pi(s')] \\
 Q^\pi(s, a) &= \mathbb{E}_{\tau \sim \rho^\pi} [r_h + \gamma r_{h+1} + \gamma^2 r_{h+2} + \dots \mid s_h = s, a_h = a] && \text{for any } h \in \mathbb{N} \\
 &= r(s, a) + \gamma \mathbb{E}_{\substack{s' \sim P(s,a) \\ a' \sim \pi(s')}} [Q^\pi(s', a')]
 \end{aligned} \tag{2.3}$$

**Exercise:** Heuristically speaking, why does it no longer matter which time step we condition on when defining the value function?

### 2.3.2 The Bellman operator is a contraction mapping

Recall from ?? that the Bellman operator  $\mathcal{J}^\pi$  for a policy  $\pi$  takes in a “value function”  $v : \mathcal{S} \rightarrow \mathbb{R}$  and returns the r.h.s. of the Bellman equation for that “value function”. In the infinite-horizon setting, this is

$$[\mathcal{J}^\pi(v)](s) := \mathbb{E}_{\substack{a \sim \pi(s) \\ s' \sim P(s,a)}} [r(s, a) + \gamma v(s')].$$

The crucial property of the Bellman operator is that it is a **contraction mapping** for any policy. Intuitively, if we start with two “value functions”  $v, u : \mathcal{S} \rightarrow \mathbb{R}$ , if we repeatedly apply the Bellman operator to each of them, they will get closer and closer together at an exponential rate.

#### Definition 2.3.1: Contraction mapping

Let  $X$  be some space with a norm  $\|\cdot\|$ . We call an operator  $f : X \rightarrow X$  a **contraction mapping** if for any  $x, y \in X$ ,

$$\|f(x) - f(y)\| \leq \gamma \|x - y\|$$

for some fixed  $\gamma \in (0, 1)$ .

**Exercise:** Show that for a contraction mapping  $f$  with coefficient  $\gamma$ , for all  $t \in \mathbb{N}$ ,

$$\|f^{(t)}(x) - f^{(t)}(y)\| \leq \gamma^t \|x - y\|,$$

i.e. that any two points will be pushed closer by at least a factor of  $\gamma$  at each iteration.

It is a powerful fact (known as the **Banach fixed-point theorem**) that every contraction mapping has a unique **fixed point**  $x^*$  such that  $f(x^*) = x^*$ . This means that if we repeatedly apply  $f$  to any starting point, we will eventually converge to  $x^*$ :

$$\|f^{(t)}(x) - x^*\| \leq \gamma^t \|x - x^*\|. \quad (2.4)$$

Let's return to the RL setting and apply this result to the Bellman operator. How can we measure the distance between two "value functions"  $v, u : \mathcal{S} \rightarrow \mathbb{R}$ ? We'll take the **supremum norm** as our distance metric:

$$\|v - u\|_\infty := \sup_{s \in \mathcal{S}} |v(s) - u(s)|,$$

i.e. we compare the "value functions" on the state that causes the biggest gap between them. Then (??) implies that if we repeatedly apply  $\mathcal{J}^\pi$  to any starting "value function", we will eventually converge to  $V^\pi$ :

$$\|(\mathcal{J}^\pi)^{(t)}(v) - V^\pi\|_\infty \leq \gamma^t \|v - V^\pi\|_\infty. \quad (2.5)$$

We'll use this useful fact to prove the convergence of several algorithms later on.

**Theorem 2.3.1: The Bellman operator is a contraction mapping**

We aim to show that

$$\|\mathcal{J}^\pi(v) - \mathcal{J}^\pi(u)\|_\infty \leq \gamma \|v - u\|_\infty.$$

**Proof:** for all states  $s \in \mathcal{S}$ ,

$$\begin{aligned} |[\mathcal{J}^\pi(v)](s) - [\mathcal{J}^\pi(u)](s)| &= \left| \mathbb{E}_{a \sim \pi(s)} \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} v(s') \right] \right. \\ &\quad \left. - \mathbb{E}_{a \sim \pi(s)} \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} u(s') \right] \right| \\ &= \gamma \left| \mathbb{E}_{s' \sim P(s, a)} [v(s') - u(s')] \right| \\ &\leq \gamma \mathbb{E}_{s' \sim P(s, a)} |v(s') - u(s')| \quad (\text{Jensen's inequality}) \\ &\leq \gamma \max_{s'} |v(s') - u(s')| \\ &= \gamma \|v - u\|_\infty. \end{aligned}$$

### 2.3.3 Tabular case (linear algebraic notation)

When there are **finitely** many states and actions, i.e.  $|\mathcal{S}|, |\mathcal{A}| < \infty$ , we call the MDP **tabular** since we can express the relevant quantities as vectors and matrices (i.e. *tables* of values):

$$r \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|} \quad P \in [0, 1]^{(|\mathcal{S} \times \mathcal{A}|) \times |\mathcal{S}|} \quad \mu \in [0, 1]^{|\mathcal{S}|}$$



$$\pi \in [0, 1]^{|\mathcal{A}| \times |\mathcal{S}|} \quad V^\pi \in \mathbb{R}^{|\mathcal{S}|} \quad Q^\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}.$$

(Verify that these types make sense!)

Note that when the policy  $\pi$  is deterministic, the actions can be determined from the states, and so we can chop off the action dimension for the rewards and state transitions:

$$\begin{aligned} r^\pi &\in \mathbb{R}^{|\mathcal{S}|} & P^\pi &\in [0, 1]^{|\mathcal{S}| \times |\mathcal{S}|} & \mu &\in [0, 1]^{|\mathcal{S}|} \\ \pi &\in \mathcal{A}^{|\mathcal{S}|} & V^\pi &\in \mathbb{R}^{|\mathcal{S}|} & Q^\pi &\in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}. \end{aligned}$$

For  $P^\pi$ , we'll treat the rows as the states and the columns as the next states. Then  $P_{s,s'}^\pi$  is the probability of transitioning from state  $s$  to state  $s'$  under policy  $\pi$ .

### Example 2.3.1: Tidying MDP

The tabular MDP from before has  $|\mathcal{S}| = 2$  and  $|\mathcal{A}| = 2$ . Let's write down the quantities for the policy  $\pi$  that tidies if and only if the room is messy:

$$r^\pi = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad P^\pi = \begin{bmatrix} 0.7 & 0.3 \\ 1 & 0 \end{bmatrix}, \quad \mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

We'll see how to evaluate this policy in the next section.

## 2.3.4 Policy evaluation

The backwards DP technique we used in the finite-horizon case (??) no longer works since there is no "final timestep" to start from. We'll need another approach to policy evaluation.

The Bellman consistency conditions yield a system of equations we can solve to evaluate a policy *exactly*. For a faster approximate solution, we can iterate the policy's Bellman operator, since we know that it has a unique fixed point at the true value function.

### Tabular case for deterministic policies

The Bellman consistency equation for a deterministic policy can be written in tabular notation as

$$V^\pi = r^\pi + \gamma P^\pi V^\pi.$$

(Unfortunately, this notation doesn't simplify the expression for  $Q^\pi$ .) This system of equations can be solved with a matrix inversion:

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi. \tag{2.6}$$

Note we've assumed that  $I - \gamma P^\pi$  is invertible. Can you see why this is the case?

(Recall that a linear operator, i.e. a square matrix, is invertible if and only if its null space is trivial; that is, it doesn't map any nonzero vector to zero. In this case, we can see that  $I - \gamma P^\pi$  is invertible because it maps any nonzero vector to a vector with at least one nonzero element.)

**Example 2.3.2: Tidying policy evaluation**

Let's use the same policy  $\pi$  that tidies if and only if the room is messy. Setting  $\gamma = 0.95$ , we must invert

$$I - \gamma P^\pi = \begin{bmatrix} 1 - 0.95 \times 0.7 & -0.95 \times 0.3 \\ -0.95 \times 1 & 1 - 0.95 \times 0 \end{bmatrix} = \begin{bmatrix} 0.335 & -0.285 \\ -0.95 & 1 \end{bmatrix}.$$

The inverse to two decimal points is

$$(I - \gamma P^\pi)^{-1} = \begin{bmatrix} 15.56 & 4.44 \\ 14.79 & 5.21 \end{bmatrix}.$$

Thus the value function is

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi = \begin{bmatrix} 15.56 & 4.44 \\ 14.79 & 5.21 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 15.56 \\ 14.79 \end{bmatrix}.$$

Let's sanity-check this result. Since rewards are at most 1, the maximum cumulative return of a trajectory is at most  $1/(1 - \gamma) = 20$ . We see that the value function is indeed slightly lower than this.

**Iterative policy evaluation**

The matrix inversion above takes roughly  $O(|\mathcal{S}|^3)$  time. Can we trade off the requirement of finding the *exact* value function for a faster *approximate* algorithm?

Let's use the Bellman operator to define an iterative algorithm for computing the value function. We'll start with an initial guess  $v^{(0)}$  with elements in  $[0, 1/(1 - \gamma)]$  and then iterate the Bellman operator:

$$v^{(t+1)} = \mathcal{J}^\pi(v^{(t)}) = r^\pi + \gamma P^\pi v^{(t)},$$

i.e.  $v^{(t)} = (\mathcal{J}^\pi)^{(t)}(v^{(0)})$ . Note that each iteration takes  $O(|\mathcal{S}|^2)$  time for the matrix-vector multiplication.

Then, as we showed in (??), by the Banach fixed-point theorem:

$$\|v^{(t)} - V^\pi\|_\infty \leq \gamma^t \|v^{(0)} - V^\pi\|_\infty.$$

How many iterations do we need for an  $\epsilon$ -accurate estimate? We can work backwards to solve for  $t$ :

$$\begin{aligned} \gamma^t \|v^{(0)} - V^\pi\|_\infty &\leq \epsilon \\ t &\geq \frac{\log(\epsilon / \|v^{(0)} - V^\pi\|_\infty)}{\log \gamma} \\ &= \frac{\log(\|v^{(0)} - V^\pi\|_\infty / \epsilon)}{\log(1/\gamma)}, \end{aligned}$$

and so the number of iterations required for an  $\epsilon$ -accurate estimate is

$$T = O\left(\frac{1}{1-\gamma} \log\left(\frac{1}{\epsilon(1-\gamma)}\right)\right). \quad (2.7)$$

Note that we've applied the inequalities  $\|v^{(0)} - V^\pi\|_\infty \leq 1/(1-\gamma)$  and  $\log(1/x) \geq 1-x$ .

### 2.3.5 Optimal policies

Now let's move on to solving for an optimal policy in the infinite-horizon case. As in the finite-horizon case (??), an **optimal policy**  $\pi^*$  is one that does at least as well as any other policy in all situations. That is, for all policies  $\pi$ , states  $s \in \mathcal{S}$ , times  $h \in \mathbb{N}$ , and initial trajectories  $\tau_h = (s_0, a_0, r_0, \dots, s_h)$  where  $s_h = s$ ,

$$\begin{aligned} V^{\pi^*}(s) &= \mathbb{E}_{\tau \sim \rho^{\pi^*}}[r_h + \gamma r_{h+1} + \gamma^2 r_{h+2} + \dots \mid s_h = s] \\ &\geq \mathbb{E}_{\tau \sim \rho^\pi}[r_h + \gamma r_{h+1} + \gamma^2 r_{h+2} + \dots \mid \tau_h] \end{aligned} \quad (2.8)$$

Once again, all optimal policies share the same **optimal value function**  $V^*$ , and the greedy policy w.r.t. this value function is optimal.

**Exercise:** Verify this by modifying the proof ?? from the finite-horizon case.

So how can we compute such an optimal policy? We can't use the backwards DP approach from the finite-horizon case (??) since there's no "final timestep" to start from. Instead, we'll exploit the fact that the Bellman consistency equation (??) for the optimal value function doesn't depend on any policy:

$$V^*(s) = \max_a \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^*(s') \right]$$

**Exercise:** Verify this by substituting the greedy policy into the Bellman consistency equation.

As before, thinking of the r.h.s. as an operator on value functions gives the **Bellman optimality operator**

$$[\mathcal{J}^*(v)](s) = \max_a \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} v(s') \right].$$

#### Value iteration

Since the optimal policy is still a policy, our result that the Bellman operator is a contracting map still holds, and so we can repeatedly apply this operator to converge to the optimal value function! This algorithm is known as **value iteration**.

**Definition 2.3.2: Value iteration pseudocode**

```

 $v^{(0)} \leftarrow 0$ 
for  $t = 0, 1, 2, \dots, T - 1$  do
   $v^{(t+1)} \leftarrow \mathcal{J}^*(v^{(t)})$ 
end for
return  $v^{(T)}$ 

```

Note that the runtime analysis for an  $\epsilon$ -optimal value function is exactly the same as iterative policy evaluation (??)! This is because value iteration is simply the special case of applying iterative policy evaluation to the *optimal* value function.

As the final step of the algorithm, to return an actual policy  $\hat{\pi}$ , we can simply act greedily w.r.t. the final iteration  $v^{(T)}$  of our above algorithm:

$$\hat{\pi}(s) = \arg \max_a \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} v^{(T)}(s') \right].$$

We must be careful, though: the value function of this greedy policy,  $V^{\hat{\pi}}$ , is *not* the same as  $v^{(T)}$ , which need not even be a well-defined value function for some policy!

The bound on the policy's quality is actually quite loose: if  $\|v^{(T)} - V^*\|_{\infty} \leq \epsilon$ , then the greedy policy  $\hat{\pi}$  satisfies  $\|V^{\hat{\pi}} - V^*\|_{\infty} \leq \frac{2\gamma}{1-\gamma}\epsilon$ , which might potentially be very large.

**Theorem 2.3.2: Greedy policy value worsening**

We aim to show that

$$\|V^{\hat{\pi}} - V^*\|_{\infty} \leq \frac{2\gamma}{1-\gamma} \|v - V^*\|_{\infty}$$

where  $\hat{\pi}(s) = \arg \max_a q(s, a)$  is the greedy policy w.r.t.

$$q(s, a) = r(s, a) + \mathbb{E}_{s' \sim P(s, a)} v(s').$$

**Proof:** We first have

$$\begin{aligned} V^*(s) - V^{\hat{\pi}}(s) &= Q^*(s, \pi^*(s)) - Q^{\hat{\pi}}(s, \hat{\pi}(s)) \\ &= [Q^*(s, \pi^*(s)) - Q^*(s, \hat{\pi}(s))] + [Q^*(s, \hat{\pi}(s)) - Q^{\hat{\pi}}(s, \hat{\pi}(s))]. \end{aligned}$$

Let's bound these two quantities separately.

For the first quantity, note that by the definition of  $\hat{\pi}$ , we have

$$q(s, \hat{\pi}(s)) \geq q(s, \pi^*(s)).$$

Let's add  $q(s, \hat{\pi}(s)) - q(s, \pi^*(s)) \geq 0$  to the first term to get

$$Q^*(s, \pi^*(s)) - Q^*(s, \hat{\pi}(s)) \leq [Q^*(s, \pi^*(s)) - q(s, \pi^*(s))] + [q(s, \hat{\pi}(s)) - Q^*(s, \hat{\pi}(s))]$$

$$\begin{aligned}
&= \gamma \mathbb{E}_{s' \sim P(s, \pi^*(s))} [V^*(s') - v(s')] + \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}(s))} [v(s') - V^*(s')] \\
&\leq 2\gamma \|v - V^*\|_\infty.
\end{aligned}$$

The second quantity is bounded by

$$\begin{aligned}
Q^*(s, \hat{\pi}(s)) - Q^{\hat{\pi}}(s, \hat{\pi}(s)) &= \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}(s))} [V^*(s') - V^{\hat{\pi}}(s')] \\
&\leq \gamma \|V^* - V^{\hat{\pi}}\|_\infty
\end{aligned}$$

and thus

$$\begin{aligned}
\|V^* - V^{\hat{\pi}}\|_\infty &\leq 2\gamma \|v - V^*\|_\infty + \gamma \|V^* - V^{\hat{\pi}}\|_\infty \\
\|V^* - V^{\hat{\pi}}\|_\infty &\leq \frac{2\gamma \|v - V^*\|_\infty}{1 - \gamma}.
\end{aligned}$$

So in order to compensate and achieve  $\|V^{\hat{\pi}} - V^*\| \leq \epsilon$ , we must have

$$\|v^{(T)} - V^*\|_\infty \leq \frac{1 - \gamma}{2\gamma} \epsilon.$$

This means, using (??), we need to run value iteration for

$$T = O\left(\frac{1}{1 - \gamma} \log\left(\frac{\gamma}{\epsilon(1 - \gamma)^2}\right)\right)$$

iterations to achieve an  $\epsilon$ -accurate estimate of the optimal value function.

### Policy iteration

Can we mitigate this “greedy worsening”? What if instead of approximating the optimal value function and then acting greedily by it at the very end, we iteratively improve the policy and value function *together*? This is the idea behind **policy iteration**. In each step, we simply set the policy to act greedily with respect to its own value function.

#### Definition 2.3.3: Policy Iteration

```

 $\pi^{(0)} : \mathcal{S} \rightarrow \mathcal{A}$  arbitrary
for  $t = 0, \dots, T - 1$  do
   $V^{\pi^{(t)}} \leftarrow (I - \gamma P^{\pi^{(t)}})^{-1} r^{\pi^{(t)}}$  ▷ (Exact) Policy Evaluation (??)
   $Q^{\pi^{(t)}}(s, a) \leftarrow r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} [V^{\pi^{(t)}}(s')]$ 
   $\pi^{(t+1)}(s) \leftarrow \arg \max_a Q^{\pi^{(t)}}(s, a)$  ▷ Policy Improvement
end for

```

Although PI appears more complex than VI, we'll use the same contraction property (??) to show convergence. This will give us the same runtime bound as value iteration and iterative

policy evaluation for an  $\epsilon$ -optimal value function (??), although in practice, PI often converges much faster.

**Theorem 2.3.3: Policy Iteration runtime and convergence**

We aim to show that the number of iterations required for an  $\epsilon$ -accurate estimate of the optimal value function is

$$T = O\left(\frac{1}{1-\gamma} \log\left(\frac{1}{\epsilon(1-\gamma)}\right)\right).$$

This bound follows from the contraction property (??):

$$\|V^{\pi^{t+1}} - V^*\|_\infty \leq \gamma \|V^{\pi^t} - V^*\|_\infty.$$

We'll prove that the iterates of PI respect the contraction property by showing that the policies improve monotonically:

$$V^{\pi^{t+1}}(s) \geq V^{\pi^t}(s).$$

Then we'll use this to show  $V^{\pi^{t+1}}(s) \geq [\mathcal{J}^*(V^{\pi^t})](s)$ . Note that

$$\begin{aligned} [\mathcal{J}^*(V^{\pi^t})](s) &= \max_a \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^{\pi^t}(s') \right] \\ &= r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} V^{\pi^t}(s') \end{aligned}$$

Since  $[\mathcal{J}^*(V^{\pi^t})](s) \geq V^{\pi^t}(s)$ , we then have

$$\begin{aligned} V^{\pi^{t+1}}(s) - V^{\pi^t}(s) &\geq V^{\pi^{t+1}}(s) - \mathcal{J}^*(V^{\pi^t})(s) \\ &= \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} \left[ V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \right]. \end{aligned} \tag{2.9}$$

But note that the expression being averaged is the same as the expression on the l.h.s. with  $s$  replaced by  $s'$ . So we can apply the same inequality recursively to get

$$\begin{aligned} V^{\pi^{t+1}}(s) - V^{\pi^t}(s) &\geq \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} \left[ V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \right] \\ &\geq \gamma^2 \mathbb{E}_{\substack{s' \sim P(s, \pi^{t+1}(s)) \\ s'' \sim P(s', \pi^{t+1}(s'))}} \left[ V^{\pi^{t+1}}(s'') - V^{\pi^t}(s'') \right] \\ &\geq \dots \end{aligned}$$

which implies that  $V^{\pi^{t+1}}(s) \geq V^{\pi^t}(s)$  for all  $s$  (since the r.h.s. converges to zero). We can then plug this back into (??) to get the desired result:

$$V^{\pi^{t+1}}(s) - \mathcal{J}^*(V^{\pi^t})(s) = \gamma \mathbb{E}_{s' \sim P(s, \pi^{t+1}(s))} \left[ V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \right]$$

$$\geq 0$$

$$V^{\pi^{t+1}}(s) \geq [\mathcal{J}^*(V^{\pi^t})](s)$$

This means we can now apply the Bellman convergence result (??) to get

$$\|V^{\pi^{t+1}} - V^*\|_\infty \leq \|\mathcal{J}^*(V^{\pi^t}) - V^*\|_\infty \leq \gamma \|V^{\pi^t} - V^*\|_\infty.$$

## 2.4 Summary

- Markov decision processes (MDPs) are a framework for sequential decision making under uncertainty. They consist of a state space  $\mathcal{S}$ , an action space  $\mathcal{A}$ , an initial state distribution  $\mu \in \Delta(\mathcal{S})$ , a transition function  $P(s' | s, a)$ , and a reward function  $r(s, a)$ . They can be finite-horizon (ends after  $H$  timesteps) or infinite-horizon (where rewards scale by  $\gamma \in (0, 1)$  at each timestep).
- Our goal is to find a policy  $\pi$  that maximizes expected total reward. Policies can be **deterministic** or **stochastic**, **state-dependent** or **history-dependent**, **stationary** or **time-dependent**.
- A policy induces a distribution over **trajectories**.
- We can evaluate a policy by computing its **value function**  $V^\pi(s)$ , which is the expected total reward starting from state  $s$  and following policy  $\pi$ . We can also compute the **state-action value function**  $Q^\pi(s, a)$ , which is the expected total reward starting from state  $s$ , taking action  $a$ , and then following policy  $\pi$ . In the finite-horizon setting, these also depend on the timestep  $h$ .
- The **Bellman consistency equation** is an equation that the value function must satisfy. It can be used to solve for the value functions exactly. Thinking of the r.h.s. of this equation as an operator on value functions gives the **Bellman operator**.
- In the finite-horizon setting, we can compute the optimal policy using **dynamic programming**.
- In the infinite-horizon setting, we can compute the optimal policy using **value iteration** or **policy iteration**.

# Contents



# Chapter 3

## Linear Quadratic Regulators

### 3.1 Motivation

Have you ever tried balancing a pen upright on your palm? If not, try it! It's a lot harder than seems. Unlike the settings we studied in the previous chapter, the state space and action space in this example aren't *finite*, or even *discrete*. Instead, they are *continuous* (real-valued) and therefore *uncountably infinite*. In addition, the state transitions governing the system – that is, the laws of physics – are highly complex.

This task is the classic CartPole *control problem*:

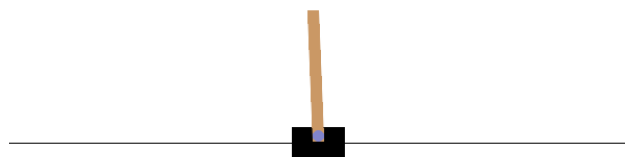


Figure 3.1: CartPole: A pole balanced on a cart.

#### Example 3.1.1: CartPole

Consider a pole balanced on a cart. The state  $x$  consists of just four continuous values:

1. The position of the cart;
2. The velocity of the cart;

3. The angle of the pole;
4. The angular velocity of the pole.

We can *control* the cart by applying a horizontal force  $u \in \mathbb{R}$ .

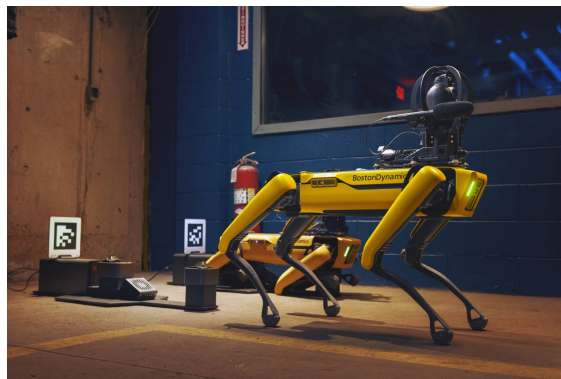
**Goal:** Stabilize the cart around an ideal state and action  $(x^*, u^*)$ .

*Controls* are the analogue to *actions* in MDPs. In control theory, the state and controls are typically denoted as  $x$  and  $u$ , but we'll stick with the  $x$  and  $u$  notation to highlight the similarity with MDPs.

Beyond this simple scenario, there are many real-world examples that involve continuous control. Here are just a few: Autonomous driving, controlling a robot's joints, and automated manufacturing. How can we teach computers to solve these kinds of problems?



(a) Solving a Rubik's Cube with a robot hand.



(b) Boston Dynamics's Spot robot.

Figure 3.2: Examples of control tasks.

In the last chapter, we developed efficient algorithms (*value iteration* and *policy iteration*) for calculating the optimal value function  $V^*$  and optimal policy  $\pi^*$  in an MDP. In this chapter, we'll develop similar algorithms for the continuous control setting.

Note that we're still assuming that the entire environment is *known* (i.e. the state transitions, rewards, etc). We'll explore the unknown case in the next chapter.

## 3.2 Optimal control

Recall that an MDP is defined by its state space  $\mathcal{S}$ , action space  $\mathcal{A}$ , state transitions  $P$ , reward function  $r$ , and discount factor  $\gamma$  or time horizon  $H$ . What are the equivalents in the control setting?

- The state and action spaces are *continuous* rather than finite. That is,  $\mathcal{S} = \mathbb{R}^{n_x}$  and  $\mathcal{A} = \mathbb{R}^{n_u}$ , where  $n_x$  and  $n_u$  are the number of coordinates to specify a single state or action respectively.

- We call the state transitions the **dynamics** of the system. In the most general case, these might change across timesteps and also include some stochastic **noise**  $w_h$ . We denote these dynamics as the function  $f_h$ , such that  $x_{h+1} = f_h(x_h, u_h, w_h)$ . Of course, we can simplify to cases where the dynamics are *deterministic/noise-free* (no  $w_h$  term) or are *stationary/time-homogeneous* (the same function  $f$  across timesteps).
- Instead of a reward function, it's more common to consider a **cost function**  $c_h : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  that describes *how far away* we are from our **goal state-action pair**  $(x^*, u^*)$ . An important special case is when the cost is *time-homogeneous*; that is, it remains the same function  $c$  at each timestep.
- We seek to minimize the *undiscounted* cost within a *time horizon*  $H$ . Note that we end an episode at  $x_H$  – there is no  $u_H$ , and so we denote the cost for the final state as  $c_H(x_H)$ .

With all of these components, we can now formulate the **optimal control problem**: *find a time-dependent policy to minimize the expected undiscounted cost over  $H$  timesteps.*

**Definition 3.2.1: Optimal control problem**

$$\begin{aligned}
 \min_{\pi_0, \dots, \pi_{H-1} : \mathcal{S} \rightarrow \mathcal{A}} \quad & \mathbb{E}_{x_0, w_h} \left[ \left( \sum_{h=0}^{H-1} c_h(x_h, u_h) \right) + c_H(x_H) \right] \\
 \text{where} \quad & x_{h+1} = f_h(x_h, u_h, w_h), \\
 & u_h = \pi_h(x_h) \\
 & x_0 \sim \mu_0 \\
 & w_h \sim \text{noise}
 \end{aligned} \tag{3.1}$$

### 3.2.1 Discretization

How does this relate to the finite horizon case? If  $x_h$  and  $u_h$  were discrete, then we'd be able to work backwards using the DP algorithms we saw before (??). As a matter of fact, let's consider what happens if we *discretize* the problem. For intuition, suppose  $n_x = n_u = 1$  (that is, states and actions are real numbers). To make  $\mathcal{S}$  and  $\mathcal{A}$  discrete, let's choose some small positive  $\epsilon$ , and simply round states and actions to the nearest multiple of  $\epsilon$ . For example, if  $\epsilon = 0.01$ , then we round  $x$  and  $u$  to two decimal spaces.

How well does this work? Even if our  $\mathcal{S}$  and  $\mathcal{A}$  are finite, the existing algorithms might take unfeasibly long to complete. Suppose our state and action spaces are bounded by some constants  $\max_{x \in \mathcal{S}} \|x\| \leq B_x$  and  $\max_{u \in \mathcal{A}} \|u\| \leq B_u$ . Then using our rounding method, we must divide *each dimension* into intervals of length  $\epsilon$ , resulting in  $(B_x/\epsilon)^{n_x}$  and  $(B_u/\epsilon)^{n_u}$  total points. To get a sense of how quickly this grows, let's consider  $\epsilon = 0.01, n_x = n_u = 10$ . Then the number of elements in our transition matrix is  $|\mathcal{S}|^2 |\mathcal{A}| = (100^{10})^2 (100^{10}) = 10^{60}$ ! Try finding a computer that'll fit that in memory! (For reference, 32 GB of memory can store  $10^9$  32-bit floating point numbers.)

So as we've seen, discretizing the problem becomes impractical as soon as the action and state spaces are even moderately high-dimensional. How can we do better?

Note that by discretizing the state and action spaces, we implicitly assumed that rounding each state or action vector by some tiny amount  $\varepsilon$  wouldn't change the behavior of the system by much; namely, that the cost and dynamics were relatively *continuous*. Can we use this continuous structure in other ways? This brings us to the topic of **Linear Quadratic Regulators**, a widely used and studied tool in control theory.

### 3.3 The Linear Quadratic Regulator

The optimal control problem stated above seems very difficult to solve. The cost function might not be convex, making optimization difficult, and the state transitions might be very complex, making it difficult to satisfy the constraints. Is there a relevant simplification that we can analyze?

We'll show that a natural structure to impose is *linear dynamics* and a (*convex*) *quadratic cost function* (in both arguments). This model is called the **linear quadratic regulator** (LQR), and is a fundamental tool in control theory.

Why are these assumptions useful? As we'll see later in the chapter, it lets us *locally approximate* more complex dynamics and cost functions using their *Taylor approximations* (up to first and second order respectively). We'll also find that even for more complex setups, we can adapt the LQR model to get surprisingly good solutions.

#### Definition 3.3.1: The linear quadratic regulator

**Linear, time-homogeneous dynamics:**

$$x_{h+1} = f(x_h, u_h, w_h) = Ax_h + Bu_h + w_h$$

**Quadratic, time-homogeneous cost function:**

$$c(x_h, u_h) = \begin{cases} x_h^\top Q x_h + u_h^\top R u_h & h < H \\ x_h^\top Q x_h & h = H \end{cases}$$

We require  $Q$  and  $R$  to both be positive definite matrices so that  $c$  has a well-defined unique minimum. We can furthermore assume without loss of generality that they are both symmetric (see exercise below).

Intuitively, the cost function punishes states and actions that are far away from the origin (i.e. both the state and action are zero vectors). More generally, we'll want to replace the origin with a *goal* state and action  $(x^*, u^*)$ . This can easily be done by replacing  $x_h$  with  $(x_h - x^*)$  and  $u_h$  with  $(u_h - u^*)$  in the expression above.

**Spherical Gaussian noise:**

$$w_h \sim \mathcal{N}(0, \sigma^2 I) \quad \forall h \in [H]$$

Putting everything together, the optimization problem we want to solve is:

$$\begin{aligned} \min_{\pi_0, \dots, \pi_{H-1}: \mathcal{S} \rightarrow \mathcal{A}} \quad & \mathbb{E} \left[ \left( \sum_{h=0}^{H-1} x_h^\top Q x_h + u_h^\top R u_h \right) + x_H^\top Q x_H \right] \\ \text{where} \quad & x_{h+1} = A x_h + B u_h + w_h \\ & u_h = \pi_h(x_h) \\ & w_h \sim \mathcal{N}(0, \sigma^2 I) \\ & x_0 \sim \mu_0. \end{aligned}$$

**Exercise:** We've set  $Q$  and  $R$  to be *symmetric* positive definite (SPD) matrices. Here we'll show that the symmetry condition can be imposed without loss of generality. Show that replacing  $Q$  with  $(Q + Q^\top)/2$  (which is symmetric) yields the same cost function.

So how do we go about analyzing this system? A good first step might be to introduce *value functions*, analogous to those from MDPs (??), to reason about the behavior of the system over the time horizon.

### Definition 3.3.2: Value functions for LQR

Given a policy  $\pi = (\pi_0, \dots, \pi_{H-1})$ , we can define the value function  $V_h^\pi : \mathcal{S} \rightarrow \mathbb{R}$  as

$$\begin{aligned} V_h^\pi(x) &= \mathbb{E} \left[ \left( \sum_{i=h}^{H-1} c(x_i, u_i) \right) + c(x_H) \mid x_h = x, \forall i \geq h. u_i = \pi_i(x_i) \right] \\ &= \mathbb{E} \left[ \left( \sum_{i=h}^{H-1} x_i^\top Q x_i + u_i^\top R u_i \right) + x_H^\top Q x_H \mid x_h = x, \forall i \geq h. u_i = \pi_i(x_i) \right] \end{aligned}$$

The expression inside the expectation is called the **cost-to-go**, since it's just the total cost starting from timestep  $h$ .

The  $Q$  function additionally conditions on the first action we take:

$$\begin{aligned} Q_h^\pi(x, u) &= \mathbb{E} \left[ \left( \sum_{i=h}^{H-1} c(x_i, u_i) \right) + c(x_H) \mid (x_h, u_h) = (x, u), \forall i \geq h. u_i = \pi_i(x_i) \right] \\ &= \mathbb{E} \left[ \left( \sum_{i=h}^{H-1} x_i^\top Q x_i + u_i^\top R u_i \right) + x_H^\top Q x_H \mid (x_h, u_h) = (x, u), \forall i \geq h. u_i = \pi_i(x_i) \right] \end{aligned}$$

As in the previous chapter, these will be instrumental in constructing the optimal policy  $\pi$  via **dynamic programming**.

## 3.4 Optimality and the Riccati Equation

In this section, we'll compute the optimal value function and policy in the LQR setting using **dynamic programming**, in a very similar way to the DP algorithms we saw in ??.

1. We'll compute  $V_H^*(x)$  as our base case.
2. Then we'll work backwards, using  $V_{h+1}^*(x)$  to compute  $Q_h^*(x, u)$ ,  $\pi_{h+1}^*(x)$ , and  $V_h^*(x)$ .

Along the way, we will prove the striking fact that  $V_h^*$  and  $\pi_h^*$  have very simple structure:  $V_h^*(x)$  is a *convex quadratic* and  $\pi_h^*(x)$  is *linear*.

### Definition 3.4.1: Optimal value functions for LQR

The **optimal value function** is the one that, at any time and in any state, achieves *minimum cost across all policies*:

$$\begin{aligned} V_h^*(x) &= \min_{\pi_h, \dots, \pi_{H-1}} V_h^\pi(x) \\ &= \min_{\pi_h, \dots, \pi_{H-1}} \mathbb{E} \left[ \left( \sum_{i=h}^{H-1} x_i^\top Q x_i + u_i^\top R u_i \right) + x_H^\top Q x_H \mid x_h = x, \forall i \geq h. u_i = \pi_i(x_i) \right] \end{aligned}$$

### Theorem 3.4.1: Optimal value function in LQR is a convex quadratic

$$V_h^*(x) = x^\top P_h x + p_h \quad \forall h \in [H]$$

for some time-dependent  $P_h \in \mathbb{R}^{n_x \times n_x}$  and  $p_h \in \mathbb{R}^{n_x}$  where  $P_h$  is SPD. Note that there is no linear term.

### Theorem 3.4.2: Optimal policy in LQR is linear

$$\pi_h^*(x) = -K_h x \quad \forall h \in [H]$$

for some  $K_h \in \mathbb{R}^{k \times d}$ . (The negative is due to convention.)

**Base case:** At the final timestep, there are no possible actions to take, and so  $V_H^*(x) = c(x) = x^\top Q x$ . Thus  $V_H^*(x) = x^\top P_H x + p_H$  where  $P_H = Q$  and  $p_H$  is the zero vector.

**Inductive hypothesis:** We seek to show that the inductive step holds for both theorems: If  $V_{h+1}^*(x)$  is a convex quadratic, then  $V_h^*(x)$  must also be a convex quadratic, and  $\pi_h^*(x)$  must be linear. We'll break this down into the following steps:

**Step 1.** Show that  $Q_h^*(x, u)$  is a convex quadratic (in both  $x$  and  $u$ ).

**Step 2.** Derive the optimal policy  $\pi_h^*(x) = \arg \min_u Q_h^*(x, u)$  and show that it's linear.

**Step 3.** Show that  $V_h^*(x)$  is a convex quadratic.

This is essentially the same proof that we wrote in the finite-horizon MDP setting, except now the state and action are *continuous* instead of finite.

We first assume the inductive hypothesis that our theorems are true at time  $h + 1$ . That is,

$$V_{h+1}^*(x) = x^\top P_{h+1}x + p_{h+1} \quad \forall x \in \mathcal{S}.$$

**Step 1.** We'll start off by demonstrating that  $Q_h^*(x)$  is a convex quadratic. Recall that the definition of  $Q_h^* : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is

$$Q_h^*(x, u) = c(x, u) + \mathbb{E}_{x' \sim f(x, u, w_{h+1})} V_{h+1}^*(x').$$

Recall  $c(x, u) = x^\top Qx + u^\top Ru$ . Let's consider the average value over the next timestep. The only randomness in the dynamics comes from the noise  $w_{h+1}$ , so we can write out this expected value as:

$$\begin{aligned} & \mathbb{E}_{x' \sim f(x, u, w_{h+1})} [V_{h+1}^*(x')] \\ = & \mathbb{E}_{w_{h+1} \sim \mathcal{N}(0, \sigma^2 I)} [V_{h+1}^*(Ax + Bu + w_{h+1})] && \text{definition of } f \\ = & \mathbb{E}_{w_{h+1}} [(Ax + Bu + w_{h+1})^\top P_{h+1} (Ax + Bu + w_{h+1}) + p_{h+1}]. && \text{inductive hypothesis} \end{aligned}$$

Summing and combining like terms, we get

$$\begin{aligned} Q_h^*(x, u) &= x^\top Qx + u^\top Ru + \mathbb{E}_{w_{h+1}} [(Ax + Bu + w_{h+1})^\top P_{h+1} (Ax + Bu + w_{h+1}) + p_{h+1}] \\ &= x^\top (Q + A^\top P_{h+1} A)x + u^\top (R + B^\top P_{h+1} B)u + 2x^\top A^\top P_{h+1} Bu \\ &\quad + \mathbb{E}_{w_{h+1}} [w_{h+1}^\top P_{h+1} w_{h+1}] + p_{h+1}. \end{aligned}$$

Note that the terms that are linear in  $w_h$  have mean zero and vanish. Now consider the remaining expectation over the noise. By expanding out the product and using linearity of expectation, we can write this out as

$$\mathbb{E}_{w_{h+1}} [w_{h+1}^\top P_{h+1} w_{h+1}] = \sum_{i=1}^d \sum_{j=1}^d (P_{h+1})_{ij} \mathbb{E}_{w_{h+1}} [(w_{h+1})_i (w_{h+1})_j].$$

When dealing with these *quadratic forms*, it's often helpful to consider the terms on the diagonal ( $i = j$ ) separately from those off the diagonal. On the diagonal, the expectation becomes

$$(P_{h+1})_{ii} \mathbb{E}(w_{h+1})_i^2 = (P_{h+1})_{ii} \text{Var}((w_{h+1})_i) = \sigma^2 (P_{h+1})_{ii}.$$

Off the diagonal, since the elements of  $w_{h+1}$  are independent, the expectation factors, and since each element has mean zero, the term disappears:

$$(P_{h+1})_{ij} \mathbb{E}(w_{h+1})_i \mathbb{E}(w_{h+1})_j = 0.$$

Thus, the only terms left are the ones on the diagonal, so the sum of these can be expressed as the trace of  $\sigma^2 P_{h+1}$ :

$$\mathbb{E}_{w_{h+1}} [w_{h+1}^\top P_{h+1} w_{h+1}] = \text{Tr}(\sigma^2 P_{h+1}).$$

Substituting this back into the expression for  $Q_h^*$ , we have:

$$Q_h^*(x, u) = x^\top (Q + A^\top P_{h+1} A) x + u^\top (R + B^\top P_{h+1} B) u + 2x^\top A^\top P_{h+1} B u + \text{Tr}(\sigma^2 P_{h+1}) + p_{h+1}. \quad (3.2)$$

As we hoped, this expression is quadratic in  $x$  and  $u$ . Furthermore, we'd like to show that it also has *positive curvature* with respect to  $u$ , i.e.

$$\nabla_{uu} Q_h^*(x, u) = R + B^\top P_{h+1} B$$

is positive definite, so that its minimum with respect to  $u$  is well-defined. This is fairly straightforward: recall that in our definition of LQR, we assumed that  $R$  is SPD (see ??). Also note that since  $P_{h+1}$  is SPD (by the inductive hypothesis), so too must be  $B^\top P_{h+1} B$ . (If this isn't clear, try proving it as an exercise!) Since the sum of two SPD matrices is also SPD, we have that  $R + B^\top P_{h+1} B$  is SPD, and so  $Q_h^*$  is indeed a convex quadratic with respect to  $u$ . A similar proof shows that it's also convex with respect to  $x$ .

**Step 2.** Now we aim to show that  $\pi_h^*$  is linear. Since  $Q_h^*$  is a convex quadratic, finding its minimum over  $u$  is easy: we simply set the gradient with respect to  $u$  equal to zero and solve for  $u$ . First, we calculate the gradient:

$$\begin{aligned} \nabla_u Q_h^*(x, u) &= \nabla_u [u^\top (R + B^\top P_{h+1} B) u + 2x^\top A^\top P_{h+1} B u] \\ &= 2(R + B^\top P_{h+1} B) u + 2(x^\top A^\top P_{h+1} B)^\top \end{aligned}$$

Setting this to zero, we get

$$\begin{aligned} 0 &= (R + B^\top P_{h+1} B) \pi_h^*(x) + B^\top P_{h+1} A x \\ \pi_h^*(x) &= (R + B^\top P_{h+1} B)^{-1} (-B^\top P_{h+1} A x) \\ &= -K_h x, \end{aligned} \quad (3.3)$$

where  $K_h = (R + B^\top P_{h+1} B)^{-1} B^\top P_{h+1} A$ .

Note that this optimal policy has an interesting property: in addition to being independent of the starting distribution  $\mu_0$  (which also happened for our finite-horizon MDP solution), it's also fully **deterministic** and isn't affected by noise! (Compare this with the discrete MDP case, where calculating our optimal policy required taking an expectation over the state transitions.)

**Step 3.** To complete our inductive proof, we must show that the inductive hypothesis is true at time  $h$ ; that is, we must prove that  $V_h^*(x)$  is a convex quadratic. Using the identity  $V_h^*(x) = Q_h^*(x, \pi^*(x))$ , we have:

$$\begin{aligned} V_h^*(x) &= Q_h^*(x, \pi^*(x)) \\ &= x^\top (Q + A^\top P_{h+1} A) x + (-K_h x)^\top (R + B^\top P_{h+1} B) (-K_h x) + 2x^\top A^\top P_{h+1} B (-K_h x) \\ &\quad + \text{Tr}(\sigma^2 P_{h+1}) + p_{h+1} \end{aligned}$$



Note that with respect to  $x$ , this is the sum of a quadratic term and a constant, which is exactly what we were aiming for!

To conclude our proof, let's concretely specify the values of  $P_h$  and  $p_h$ . The constant term is clearly  $p_h = \text{Tr}(\sigma^2 P_{h+1}) + p_{h+1}$ . We can simplify the quadratic term by substituting in  $K_h$ . Notice that when we do this, the  $(R + B^\top P_{h+1} B)$  term in the expression is cancelled out by its inverse, and the remaining terms combine to give what is known as the **Riccati equation**:

**Definition 3.4.2: Riccati equation**

$$P_h = Q + A^\top P_{h+1} A - A^\top P_{h+1} B (R + B^\top P_{h+1} B)^{-1} B^\top P_{h+1} A.$$

There are several nice properties to note about the Riccati equation:

1. It's defined **recursively**. Given the dynamics defined by  $A$  and  $B$ , and the state cost matrix  $Q$ , we can recursively calculate  $P_h$  across all timesteps starting from  $P_H = Q$ .
2.  $P_h$  often appears in calculations surrounding optimality, such as  $V_h^*$ ,  $Q_h^*$ , and  $\pi_h^*$ .
3. Together with the dynamics given by  $A$  and  $B$ , and the action coefficients  $R$ , it fully defines the optimal policy.

Now we've shown that  $V_h^*(x) = x^\top P_h x + p_h$ , which is a convex quadratic, and this concludes our proof. ■

In summary, we just demonstrated that:

- The optimal value function  $V_h^*$  is convex at all  $h$ .
- The optimal  $Q$ -function  $Q_h^*$  is convex (in both arguments) at all  $h$ .
- The optimal policy  $\pi_h^*$  is linear at all  $h$ .
- All of these quantities can be calculated using a symmetric matrix  $P_h$  for each timestep, which can be defined recursively using the Riccati equation.

Before we move on to some extensions of LQR, let's consider how the state at time  $h$  behaves when we act according to this optimal policy.

### 3.4.1 Expected state at time $h$

How can we compute the expected state at time  $h$  when acting according to the optimal policy? Let's first express  $x_h$  in a cleaner way in terms of the history. Note that having linear dynamics makes it easy to expand terms backwards in time:

$$\begin{aligned} x_h &= Ax_{h-1} + Bu_{h-1} + w_{h-1} \\ &= A(Ax_{h-2} + Bu_{h-2} + w_{h-2}) + Bu_{h-1} + w_{h-1} \\ &= \dots \\ &= A^h x_0 + \sum_{i=0}^{h-1} A^i (Bu_{h-i-1} + w_{h-i-1}). \end{aligned}$$

Let's consider the *average state* at this time, given all the past states and actions. Since we assume that  $\mathbb{E}[w_h] = 0$  (this is the zero vector in  $d$  dimensions), when we take an expectation, the  $w_h$  term vanishes due to linearity, and so we're left with

$$\mathbb{E}[x_h \mid x_{0:(h-1)}, u_{0:(h-1)}] = A^h x_0 + \sum_{i=0}^{h-1} A^i B u_{h-i-1}.$$

If we choose actions according to our optimal policy, this becomes

$$\mathbb{E}[x_h \mid x_0, \forall i \leq h. u_i = -K_i x_i] = \left( \prod_{i=0}^{h-1} (A - BK_i) \right) x_0.$$

**Exercise:** Verify this.

This introduces the quantity  $A - BK_i$ , which shows up frequently in control theory. For example, one important question is: will  $x_h$  remain bounded, or will it go to infinity as time goes on? To answer this, let's imagine that these  $K_i$ s are equal (call this matrix  $K$ ). Then the expression above becomes  $(A - BK)^h x_0$ . Now consider the maximum eigenvalue  $\lambda_{\max}$  of  $A - BK$ . If  $|\lambda_{\max}| > 1$ , then there's some nonzero initial state  $\bar{x}_0$ , the corresponding eigenvector, for which

$$\lim_{h \rightarrow \infty} (A - BK)^h \bar{x}_0 = \lim_{h \rightarrow \infty} \lambda_{\max}^h \bar{x}_0 = \infty.$$

Otherwise, if  $|\lambda_{\max}| < 1$ , then it's impossible for your original state to explode as dramatically.

## 3.5 Extensions

We've now formulated an optimal solution for the time-homogeneous LQR and computed the expected state under the optimal policy. However, real world tasks rarely have such simple dynamics, and we may wish to design more complex cost functions. In this section, we'll consider more general extensions of LQR where some of the assumptions we made above are relaxed. Specifically, we'll consider:

1. **Time-dependency**, where the dynamics and cost function might change depending on the timestep.
2. **General quadratic cost**, where we allow for linear terms and a constant term.
3. **Tracking a goal trajectory** rather than aiming for a single goal state-action pair.

### 3.5.1 Time-dependent dynamics and cost function

So far, we've considered the *time-homogeneous* case, where the dynamics and cost function stay the same at every timestep. However, this might not always be the case. For example, if we want to preserve the temperature in a greenhouse, the outside forces are going to change depending on the time of day. As another example, in many sports or video games, the rules and scoring system might change during overtime. To address these sorts of problems, we can loosen the time-homogeneous restriction, and consider the case where the dynamics and cost function are

*time-dependent.* Our analysis remains almost identical; in fact, we can simply add a time index to the matrices  $A$  and  $B$  that determine the dynamics and the matrices  $Q$  and  $R$  that determine the cost. (As an exercise, walk through the derivation and verify this claim!)

The modified problem is now defined as follows:

**Definition 3.5.1: Time-dependent LQR**

$$\arg \min_{\pi_0, \dots, \pi_{H-1}} \mathbb{E} \left[ \left( \sum_{h=0}^{H-1} (x_h^\top Q_h x_h + u_h^\top R_h u_h) \right) + x_H^\top Q_H x_H \right]$$

where  $x_{h+1} = f_h(x_h, u_h, w_h) = A_h x_h + B_h u_h + w_h$   
 $x_0 \sim \mu_0$   
 $u_h = \pi_h(x_h)$   
 $w_h \sim \mathcal{N}(0, \sigma^2 I).$

The derivation of the optimal value functions and the optimal policy remains almost exactly the same, and we can modify the Riccati equation accordingly:

**Definition 3.5.2: Time-dependent Riccati Equation**

$$P_h = Q_h + A_h^\top P_{h+1} A_h - A_h^\top P_{h+1} B_h (R_h + B_h^\top P_{h+1} B_h)^{-1} B_h^\top P_{h+1} A_h.$$

Note that this is just the time-homogeneous Riccati equation (??), but with the time index added to each of the relevant matrices.

Additionally, by allowing the dynamics to vary across time, we gain the ability to *locally approximate* nonlinear dynamics at each timestep. We'll discuss this later in the chapter.

### 3.5.2 More general quadratic cost functions

Our original cost function had only second-order terms with respect to the state and action. We can also consider more general quadratic cost functions that also have first-order terms and a constant term. Combining this with time-dependent dynamics results in the following expression, where we introduce a new matrix  $M_h$  for the cross term, linear coefficients  $q_h$  and  $r_h$  for the state and action respectively, and a constant term  $c_h$ :

$$c_h(x_h, u_h) = (x_h^\top Q_h x_h + x_h^\top M_h u_h + u_h^\top R_h u_h) + (x_h^\top q_h + u_h^\top r_h) + c_h. \quad (3.4)$$

Similarly, we can also include a constant term  $v_h \in \mathbb{R}^{n_x}$  in the dynamics (note that this is *deterministic* at each timestep, unlike the stochastic noise  $w_h$ ):

$$x_{h+1} = f_h(x_h, u_h, w_h) = A_h x_h + B_h u_h + v_h + w_h.$$

The derivation of the optimal solution in this case will be left as a homework exercise.

### 3.5.3 Tracking a predefined trajectory

So far, we've been trying to get the robot to stay as close as possible to the origin, or more generally a goal state-action pair  $(x^*, u^*)$ . However, consider applying LQR to a task like autonomous driving, where the desired state changes over time, instead of remaining in one location. In these cases, we want the robot to follow a predefined *trajectory* of states and actions  $(x_h^*, u_h^*)_{h=0}^{H-1}$ . To express this as a control problem, we'll need a corresponding time-dependent cost function:

$$c_h(x_h, u_h) = (x_h - x_h^*)^\top Q(x_h - x_h^*) + (u_h - u_h^*)^\top R(u_h - u_h^*).$$

Note that this punishes states and actions that are far from the intended trajectory. By expanding out these multiplications, we can see that this is actually a special case of the more general quadratic cost function we discussed above (??):

$$M_h = 0, \quad q_h = -2Qx_h^*, \quad r_h = -2Ru_h^*, \quad c_h = (x_h^*)^\top Q(x_h^*) + (u_h^*)^\top R(u_h^*).$$

## 3.6 Approximating nonlinear dynamics

The LQR algorithm solves for the optimal policy when the dynamics are *linear* and the cost function is a *convex quadratic*. However, real settings are rarely this simple! Let's return to the CartPole example from the start of the chapter (??). The dynamics (physics) aren't linear, and we might also want to specify a cost function that's more complex.

Concretely, let's consider a *noise-free* problem since, as we saw, the noise doesn't factor into the optimal policy. Let's assume the dynamics and cost function are stationary, and ignore the terminal state for simplicity:

#### Definition 3.6.1: Nonlinear control problem

$$\begin{aligned} \min_{\pi_0, \dots, \pi_{H-1}: \mathcal{S} \rightarrow \mathcal{A}} \quad & \mathbb{E}_{x_0} \left[ \sum_{h=0}^{H-1} c(x_h, u_h) \right] \\ \text{where} \quad & x_{h+1} = f(x_h, u_h) \\ & u_h = \pi_h(x_h) \\ & x_0 \sim \mu_0 \\ & c(x, u) = d(x, x^*) + d(u, u^*). \end{aligned}$$

Here,  $d$  denotes some general distance metric between its two arguments.

This is now only slightly simplified from the general optimal control problem (see ??). Here, we don't know an analytical form for the dynamics  $f$  or the cost function  $c$ , but we assume that we're able to *query/sample/simulate* them to get their values at a given state and action. To clarify, consider the case where the dynamics are given by real world physics. We can't (yet) write down an expression for the dynamics that we can differentiate or integrate analytically. However, we can still *simulate* the dynamics and cost function by running a real-world experiment and measuring the resulting states and costs. How can we adapt LQR to this more general nonlinear case?

### 3.6.1 Local linearization

How can we apply LQR when the dynamics are nonlinear or the cost function is more complex? We'll exploit the useful fact that we can take any *locally continuous* function and approximate it using a Taylor expansion of low-order polynomials. In particular, as long as the dynamics  $f$  are differentiable around  $(x^*, u^*)$  and the cost function  $c$  is twice differentiable at  $(x^*, u^*)$ , we can take a linear approximation of  $f$  and a quadratic approximation of  $c$  to bring us back to the regime of LQR.

Linearizing the dynamics around  $(x^*, u^*)$  gives:

$$f(x, u) \approx f(x^*, u^*) + \nabla_x f(x^*, u^*)(x - x^*) + \nabla_u f(x^*, u^*)(u - u^*)$$

$$(\nabla_x f(x, u))_{ij} = \frac{df_i(x, u)}{dx_j}, \quad i, j \leq n_x \quad (\nabla_u f(x, u))_{ij} = \frac{df_i(x, u)}{du_j}, \quad i \leq n_x, j \leq n_u$$

and quadratizing the cost function around  $(x^*, u^*)$  gives:

$$\begin{aligned} c(x, u) &\approx c(x^*, u^*) && \text{constant term} \\ &+ \nabla_x c(x^*, u^*)(x - x^*) + \nabla_u c(x^*, u^*)(u - u^*) && \text{linear terms} \\ &+ \frac{1}{2}(x - x^*)^\top \nabla_{xx} c(x^*, u^*)(x - x^*) && \text{quadratic terms} \\ &+ \frac{1}{2}(u - u^*)^\top \nabla_{uu} c(x^*, u^*)(u - u^*) \\ &+ (x - x^*)^\top \nabla_{xu} c(x^*, u^*)(u - u^*) \end{aligned}$$

where the gradients and Hessians are defined as

$$\begin{aligned} (\nabla_x c(x, u))_i &= \frac{dc(x, u)}{dx_i}, \quad i \leq n_x & (\nabla_u c(x, u))_i &= \frac{dc(x, u)}{du_i}, \quad i \leq n_u \\ (\nabla_{xx} c(x, u))_{ij} &= \frac{d^2 c(x, u)}{dx_i dx_j}, \quad i, j \leq n_x & (\nabla_{uu} c(x, u))_{ij} &= \frac{d^2 c(x, u)}{du_i du_j}, \quad i, j \leq n_u \\ (\nabla_{xu} c(x, u))_{ij} &= \frac{d^2 c(x, u)}{dx_i du_j}, \quad i \leq n_x, j \leq n_u \end{aligned}$$

**Exercise:** Note that this cost can be expressed in the general quadratic form seen in ???. Derive the corresponding quantities  $Q, R, M, q, r, c$ .

### 3.6.2 Finite differencing

To calculate these gradients and Hessians in practice, we use a method known as **finite differencing** for numerically computing derivatives. Namely, we can simply use the limit definition of the derivative, and see how the function changes as we add or subtract a tiny  $\delta$  to the input. Note that this only requires us to be able to *query* the function, not to have an analytical expression for it, which is why it's so useful in practice.

### 3.6.3 Local convexification

However, simply taking the second-order approximation of the cost function is insufficient, since for LQR we required that the  $Q$  and  $R$  matrices were positive definite, i.e. that all of their eigenvalues were positive.

One way to naively *force* some symmetric matrix  $D$  to be positive definite is to set any non-positive eigenvalues to some small positive value  $\varepsilon > 0$ . Recall that any real symmetric matrix  $D \in \mathbb{R}^{n \times n}$  has an basis of eigenvectors  $u_1, \dots, u_n$  with corresponding eigenvalues  $\lambda_1, \dots, \lambda_n$  such that  $Du_i = \lambda_i u_i$ . Then we can construct the positive definite approximation by

$$\tilde{D} = \sum_{i=1, \dots, n | \lambda_i > 0} \lambda_i u_i u_i^\top + \varepsilon I.$$

**Exercise:** Convince yourself that  $\tilde{D}$  is indeed positive definite.

Note that Hessian matrices are generally symmetric, so we can apply this process to  $Q$  and  $R$  to obtain the positive definite approximations  $\tilde{Q}$  and  $\tilde{R}$ . Now that we have a convex quadratic approximation to the cost function, and a linear approximation to the state transitions, we can simply apply the time-homogenous LQR methods from ??.

But what happens when we enter states far away from  $x^*$  or want to use actions far from  $u^*$ ? A Taylor approximation is only accurate in a *local* region around the point of linearization, so the performance of our LQR controller will degrade as we move further away. We'll see how to address this in the next section using the **iterative LQR** algorithm.

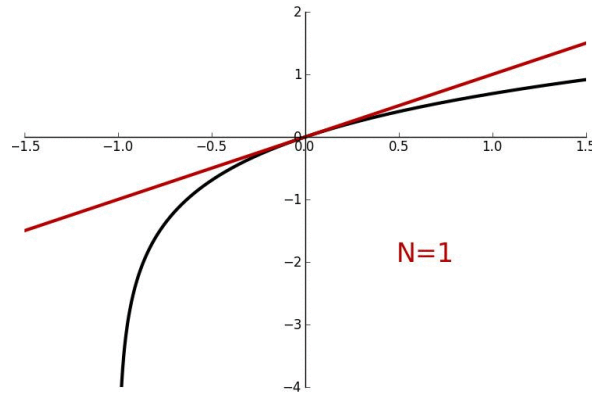


Figure 3.3: A visual intuition for local linearization.

### 3.6.4 Iterative LQR

Earlier, we tried to use local linearization to solve nonlinear control problems with the LQR, but we ran into the issue that the approximation is only accurate in a local region around the point of linearization. Instead, we'll use an iterative approach, where we repeatedly linearize around

different points to create a *time-dependent* approximation of the dynamics, and then solve the resulting time-dependent LQR problem to obtain a better policy. This is known as **iterative LQR** or **iLQR**:

**Definition 3.6.2: Iterative LQR (high-level)**

For each iteration of the algorithm:

- Step 1.** Form a time-dependent LQR problem around the current candidate trajectory using local linearization.
- Step 2.** Compute the optimal policy using ??.
- Step 3.** Generate a new series of actions using this policy.
- Step 4.** Compute a better candidate trajectory by interpolating between the current and proposed actions.

Now let's go through the details of each step. We'll use superscripts to denote the iteration of the algorithm. We'll also denote  $\bar{x}_0 = \mathbb{E}_{x_0 \sim \mu_0}[x_0]$  as the expected initial state.

At iteration  $i$  of the algorithm, we begin with a **candidate** trajectory  $\bar{\tau}^i = (\bar{x}_0^i, \bar{u}_0^i, \dots, \bar{x}_{H-1}^i, \bar{u}_{H-1}^i)$ .

**Step 1: Form a time-dependent LQR problem.** At each timestep  $h \in [H]$ , we use the techniques from ?? to linearize the dynamics and quadratize the cost function around  $(\bar{x}_h^i, \bar{u}_h^i)$ :

$$\begin{aligned} f_h(x, u) &\approx f(\bar{x}_h^i, \bar{u}_h^i) + \nabla_x f(\bar{x}_h^i, \bar{u}_h^i)(x - \bar{x}_h^i) + \nabla_u f(\bar{x}_h^i, \bar{u}_h^i)(u - \bar{u}_h^i) \\ c_h(x, u) &\approx c(\bar{x}_h^i, \bar{u}_h^i) + \begin{bmatrix} x - \bar{x}_h^i & u - \bar{u}_h^i \end{bmatrix} \begin{bmatrix} \nabla_x c(\bar{x}_h^i, \bar{u}_h^i) \\ \nabla_u c(\bar{x}_h^i, \bar{u}_h^i) \end{bmatrix} \\ &\quad + \frac{1}{2} \begin{bmatrix} x - \bar{x}_h^i & u - \bar{u}_h^i \end{bmatrix} \begin{bmatrix} \nabla_{xx} c(\bar{x}_h^i, \bar{u}_h^i) & \nabla_{xu} c(\bar{x}_h^i, \bar{u}_h^i) \\ \nabla_{ux} c(\bar{x}_h^i, \bar{u}_h^i) & \nabla_{uu} c(\bar{x}_h^i, \bar{u}_h^i) \end{bmatrix} \begin{bmatrix} x - \bar{x}_h^i \\ u - \bar{u}_h^i \end{bmatrix}. \end{aligned}$$

**Step 2: Compute the optimal policy.** We can now solve the time-dependent LQR problem using the Riccati equation from ?? to compute the optimal policy  $\pi_0^i, \dots, \pi_{H-1}^i$ .

**Step 3: Generate a new series of actions.** We can then generate a new sample trajectory by taking actions according to this optimal policy:

$$\bar{x}_0^{i+1} = \bar{x}_0, \quad \tilde{u}_h = \pi_h^i(\bar{x}_h^{i+1}), \quad \bar{x}_{h+1}^{i+1} = f(\bar{x}_h^{i+1}, \tilde{u}_h).$$

Note that the states are sampled according to the *true* dynamics, which we assume we have query access to.

**Step 4: Compute a better candidate trajectory.** Note that we've denoted these actions as  $\tilde{u}_h$  and aren't directly using them for the next iteration  $\bar{u}_h^{i+1}$ . Rather, we want to *interpolate* between them and the actions from the previous iteration  $\bar{u}_0^i, \dots, \bar{u}_{H-1}^i$ . This is so that the cost will *increase monotonically*, since if the new policy turns out to actually be worse, we can stay closer to the previous trajectory. (Can you think of an intuitive example where this might happen?)

Formally, we want to find  $\alpha \in [0, 1]$  to generate the next iteration of actions  $\bar{u}_0^{i+1}, \dots, \bar{u}_{H-1}^{i+1}$  such that the cost is minimized:

$$\begin{aligned} \min_{\alpha \in [0, 1]} \quad & \sum_{h=0}^{H-1} c(x_h, \bar{u}_h^{i+1}) \\ \text{where} \quad & x_{h+1} = f(x_h, \bar{u}_h^{i+1}) \\ & \bar{u}_h^{i+1} = \alpha \bar{u}_h^i + (1 - \alpha) \tilde{u}_h \\ & x_0 = \bar{x}_0. \end{aligned}$$

Note that this optimizes over the closed interval  $[0, 1]$ , so by the Extreme Value Theorem, it's guaranteed to have a global maximum.

The final output of this algorithm is a policy  $\pi^{n_{\text{steps}}}$  derived after  $n_{\text{steps}}$  of the algorithm. Though the proof is somewhat complex, one can show that for many nonlinear control problems, this solution converges to a locally optimal solution (in the policy space).



# Chapter 4

## Policy Gradients

### 4.1 Motivation

The scope of our problem has been gradually expanding.

1. In the first chapter, we considered *bandits* with a finite number of arms, where the only stochasticity involved was their rewards.
2. In the second chapter, we considered *MDPs* more generally, involving a finite number of states and actions, where the state transitions are Markovian.
3. In the third chapter, we considered *continuous* state and action spaces and developed the *Linear Quadratic Regulator*. We then showed how to use it to find *locally optimal solutions* to problems with nonlinear dynamics and non-quadratic cost functions.

Now, we'll continue to investigate the case of finding optimal policies in large MDPs using the self-explanatory approach of *policy optimization*. This is a general term encompassing many specific algorithms we've already seen:

- *Policy iteration* for finite MDPs,
- *Iterative LQR* for locally optimal policies in continuous control.

Here we'll see some general algorithms that allow us to optimize policies for general kinds of problems. These algorithms have been used in many groundbreaking applications, including AlphaGo, OpenAI Five. These methods also bring us into the domain where we can use *deep learning* to approximate complex, nonlinear functions.

### 4.2 (Stochastic) Policy Gradient Ascent

Let's suppose our policy can be *parameterized* by some parameters  $\theta$ . For example, these might be a preferences over state-action pairs, or in a high-dimensional case, the weights and biases of a deep neural network. We'll talk more about possible parameterizations in ??

Remember that in reinforcement learning, the goal is to *maximize reward*. Specifically, we seek the parameters that maximize the expected total reward, which we can express concisely using the value function we defined earlier:

$$\begin{aligned}
 J(\theta) &:= \mathbb{E}_{s_0 \sim \mu_0} V^{\pi_\theta}(s_0) = \mathbb{E} \sum_{t=0}^{T-1} r_t \\
 &\quad \text{where } s_0 \sim \mu_0 \\
 &\quad s_{t+1} \sim P(s_t, a_t), \\
 &\quad a_h = \pi_\theta(s_h) \\
 &\quad r_h = r(s_h, a_h).
 \end{aligned} \tag{4.1}$$

We call a sequence of states, actions, and rewards a **trajectory**  $\tau = (s_i, a_i, r_i)_{i=0}^{T-1}$ , and the total time-discounted reward is also often called the **return**  $R(\tau)$  of a trajectory. Note that the above is the *undiscounted, finite-horizon case*, which we'll continue to use throughout the chapter, but analogous results hold for the *discounted, infinite-horizon case*.

Note that when the state transitions are Markov (i.e.  $s_t$  only depends on  $s_{t-1}, a_{t-1}$ ) and the policy is stationary (i.e.  $a_t \sim \pi_\theta(s_t)$ ), we can write out the *likelihood of a trajectory* under the policy  $\pi_\theta$ :

$$\begin{aligned}
 \rho_\theta(\tau) &= \mu(s_0) \pi_\theta(a_0|s_0) \\
 &\quad \times P(s_1|s_0, a_0) \pi_\theta(a_1|s_1) \\
 &\quad \times \dots \\
 &\quad \times P(s_{H-1}|s_{H-2}, a_{H-2}) \pi_\theta(a_{H-1}|s_{H-1}).
 \end{aligned} \tag{4.2}$$

This lets us rewrite  $J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta} R(\tau)$ .

Now how do we optimize for this function (the expected total reward)? One very general optimization technique is *gradient ascent*. Namely, the **gradient** of a function at a given point answers: At this point, which direction should we move to increase the function the most? By repeatedly moving in this direction, we can keep moving up on the graph of this function. Expressing this iteratively, we have:

$$\theta_{t+1} = \theta_t + \eta \nabla_{\theta} J(\pi_\theta) \Big|_{\theta=\theta_t},$$

Where  $\eta$  is a *hyperparameter* that says how big of a step to take each time.

In order to apply this technique, we need to be able to evaluate the gradient  $\nabla_{\theta} J(\pi_\theta)$ . How can we do this?

In practice, it's often impractical to evaluate the gradient directly. For example, in supervised learning,  $J(\theta)$  might be the sum of squared prediction errors across an entire **training dataset**. However, if our dataset is very large, we might not be able to fit it into our computer's memory!

Instead, we can *estimate* a gradient step using some estimator  $\tilde{\nabla} J(\theta)$ . This is called **stochastic gradient descent** (SGD). Ideally, we want this estimator to be **unbiased**, that is, on average,

it matches a single true gradient step:

$$\mathbb{E}[\tilde{\nabla} J(\theta)] = \nabla J(\theta).$$

If  $J$  is defined in terms of some training dataset, we might randomly choose a *minibatch* of samples and use them to estimate the prediction error across the *whole* dataset. (This approach is known as **minibatch SGD**.)

Notice that our parameters will stop changing once  $\nabla J(\theta) = 0$ . This implies that our current parameters are ‘locally optimal’ in some sense; it’s impossible to increase the function by moving in any direction. If  $J$  is convex, then the only point where this happens is at the *global optimum*. Otherwise, if  $J$  is nonconvex, the best we can hope for is a *local optimum*.

We can actually show that in a finite number of steps, SGD will find a  $\theta$  that is “close” to a local optimum. More formally, suppose we run SGD for  $T$  steps, using an unbiased gradient estimator. Let the step size  $\eta_t$  scale as  $O(1/\sqrt{t})$ . Then if  $J$  is bounded and  $\beta$ -smooth, and the norm of the gradient estimator has a finite variance, then after  $T$  steps:

$$\|\nabla_{\theta} J(\theta)\|^2 \leq O(M\beta\sigma^2/T).$$

In another perspective, the local “landscape” of  $J$  around  $\theta$  becomes flatter and flatter the longer we run SGD.

## 4.3 REINFORCE and Importance Sampling

Note that the objective function above,  $J(\theta) = \mathbb{E}_{\tau \sim \rho_{\theta}} R(\tau)$ , is very difficult, or even intractable, to compute exactly! This is because it involves taking an expectation over all possible trajectories  $\tau$ . Can we rewrite this in a form that’s more convenient to implement?

Specifically, suppose there is some distribution over trajectories  $\rho(\tau)$  that’s easy to sample from (e.g. a database of existing trajectories). We can then rewrite the gradient of objective function, a.k.a. the *policy gradient*, as follows (all gradients are being taken w.r.t.  $\theta$ ):

$$\begin{aligned} \nabla J(\theta) &= \nabla \mathbb{E}_{\tau \sim \rho_{\theta}} [R(\tau)] \\ &= \nabla \mathbb{E}_{\tau \sim \rho} \left[ \frac{\rho_{\theta}(\tau)}{\rho(\tau)} R(\tau) \right] && \text{likelihood ratio trick} \\ &= \mathbb{E}_{\tau \sim \rho} \left[ \frac{\nabla \rho_{\theta}(\tau)}{\rho(\tau)} R(\tau) \right] && \text{switching gradient and expectation} \end{aligned}$$

Note that setting  $\rho = \rho_{\theta}$  allows us to express  $\nabla J$  as an expectation. (Notice the swapped order of  $\nabla$  and  $\mathbb{E}$ !)

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim \rho_{\theta}} [\nabla \log \rho_{\theta}(\tau) \cdot R(\tau)].$$

Consider expanding out  $\rho_{\theta}$ . Note that taking its log turns it into a sum of log terms, of which only the  $\pi_{\theta}(a_t|s_t)$  terms depend on  $\theta$ , so we can simplify even further to obtain the following

expression for the policy gradient, known as the “REINFORCE” policy gradient:

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta} \left[ \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t | s_t) R(\tau) \right]$$

This expression allows us to estimate the gradient by sampling a few sample trajectories from  $\pi_\theta$ , calculating the likelihoods of the chosen actions, and substituting these into the expression above.

In fact, we can perform one more simplification. Intuitively, the action taken at step  $t$  does not affect the reward from previous timesteps, since they’re already in the past! You can also show rigorously that this is the case, and that we only need to consider the present and future rewards to calculate the policy gradient:

$$\begin{aligned} \nabla J(\theta) &= \mathbb{E}_{\tau \sim \rho_\theta} \left[ \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t | s_t) \sum_{t'=t}^{T-1} r(s_{t'}, a_{t'}) \right] \\ &= \mathbb{E}_{\tau \sim \rho_\theta} \left[ \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t | s_t) Q^{\pi_\theta}(s_t, a_t) \right] \end{aligned} \quad (4.3)$$

**Exercise:** Prove that this is equivalent to the previous definitions. What modification to the expression must be made for the discounted, infinite-horizon setting?

For some intuition into how this method works, recall that we update our parameters according to

$$\begin{aligned} \theta_{t+1} &= \theta_t + \nabla J(\theta_t) \\ &= \theta_t + \mathbb{E}_{\tau \sim \rho_{\theta_t}} \nabla \log \rho_{\theta_t}(\tau) \cdot R(\tau). \end{aligned}$$

Consider the “good” trajectories where  $R(\tau)$  is large. Then  $\theta$  gets updated so that these trajectories become more likely. To see why, recall that  $\rho_\theta(\tau)$  is the likelihood of the trajectory  $\tau$  under the policy  $\pi_\theta$ , so evaluating the gradient points in the direction that makes  $\tau$  more likely.

This is an example of **importance sampling**: updating a distribution to put more density on “more important” samples (in this case trajectories).

## 4.4 Baselines and advantages

A central idea from supervised learning is the bias-variance tradeoff. So far, our method is *unbiased*, meaning that its average is the true policy gradient. Can we find ways to reduce the variance of our estimator as well?

We can instead subtract a **baseline function**  $b_t : \mathcal{S} \rightarrow \mathbb{R}$  at each timestep  $t$ . This modifies the policy gradient as follows:

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta} \left[ \sum_{t=0}^{H-1} \nabla \log \pi_\theta(a_t | s_t) \left( \left( \sum_{t'=t}^{H-1} r_{t'} \right) - b_t(s_t) \right) \right]. \quad (4.4)$$

For example, we might want  $b_t$  to estimate the average reward-to-go at a given timestep:  $b_t^\theta = \mathbb{E}_{\tau \sim \rho_\theta} R_t(\tau)$ . This way, the random variable  $R_t(\tau) - b_t^\theta$  is centered around zero, making certain algorithms more stable.

As a better baseline, we could instead choose the *value function*. Note that the random variable  $Q_t^\pi(s, a) - V_t^\pi(s)$ , where the randomness is taken over the actions, is also centered around zero. (Recall  $V_t^\pi(s) = \mathbb{E}_{a \sim \pi} Q_t^\pi(s, a)$ .) In fact, this quantity has a particular name: the **advantage function**. This measures how much better this action does than the average for that policy. (Note that for an optimal policy  $\pi^*$ , the advantage of a given state-action pair is always nonpositive.)

We can now express the policy gradient as follows. Note that the advantage function effectively replaces the  $Q$ -function from ??:

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta} \left[ \sum_{t=0}^{T-1} \nabla \log \pi_\theta(a_t | s_t) A_t^{\pi_\theta}(s_t, a_t) \right]. \quad (4.5)$$

Note that to avoid correlations between the gradient estimator and the value estimator (i.e. baseline), we must estimate them with independently sampled trajectories:

#### Definition 4.4.1: Policy gradient with a learned baseline

**Require:** Learning rate  $\eta_0, \dots, \eta_{K-1}$

**Require:** Initialization  $\theta^0$

**for**  $k = 0, \dots, K - 1$  **do**

Sample  $N$  trajectories from  $\pi_{\theta^k}$  to estimate a baseline  $\tilde{b}$  such that  $\tilde{b}_h(s) \approx V_h^{\theta^k}(s)$

Sample  $M$  trajectories  $\tau_0, \dots, \tau_{M-1} \sim \rho_{\theta^k}$

Compute the policy gradient estimate

$$\tilde{\nabla}_\theta J(\theta^k) = \frac{1}{M} \sum_{m=0}^{M-1} \sum_{h=0}^{H-1} \nabla \log \pi_{\theta^k}(a_h | s_h) (R_h(\tau_m) - \tilde{b}_h(s_h))$$

Gradient ascent update  $\theta^{k+1} \leftarrow \theta^k + \tilde{\nabla}_\theta J(\theta^k)$

**end for**

The baseline estimation step can be done using any appropriate supervised learning algorithm. Note that the gradient estimator will be unbiased regardless of the baseline.

## 4.5 Example policy parameterizations

What are some different ways we could parameterize our policy?

If both the state and action spaces are finite, perhaps we could simply learn a preference value  $\theta_{s,a}$  for each state-action pair. Then to turn this into a valid distribution, we perform a “softmax”

operation: we exponentiate each of them, and divide by the total:

$$\pi_{\theta}^{\text{softmax}}(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{s,a'} \exp(\theta_{s,a'})}.$$

However, this doesn't make use of any structure in the states or actions, so while this is flexible, it is also prone to overfitting.

### 4.5.1 Linear in features

Instead, what if we map each state-action pair into some **feature space**  $\phi(s, a) \in \mathbb{R}^p$ ? Then, to map a feature vector to a probability, we take a linear combination  $\theta \in \mathbb{R}^p$  of the features and take a softmax:

$$\pi_{\theta}^{\text{linear in features}}(a|s) = \frac{\exp(\theta^{\top} \phi(s, a))}{\sum_{a'} \exp(\theta^{\top} \phi(s, a'))}.$$

Another interpretation is that  $\theta$  represents the feature vector of the “ideal” state-action pair, as state-action pairs whose features align closely with  $\theta$  are given higher probability.

The score function for this parameterization is also quite elegant:

$$\begin{aligned} \nabla \log \pi_{\theta}(a|s) &= \nabla \left( \theta^{\top} \phi(s, a) - \log \left( \sum_{a'} \exp(\theta^{\top} \phi(s, a')) \right) \right) \\ &= \phi(s, a) - \mathbb{E}_{a' \sim \pi_{\theta}(s)} \phi(s, a') \end{aligned}$$

Plugging this into our policy gradient expression, we get

$$\begin{aligned} \nabla J(\theta) &= \mathbb{E}_{\tau \sim \rho_{\theta}} \left[ \sum_{t=0}^{T-1} \nabla \log \pi_{\theta}(a_t|s_t) A_t^{\pi_{\theta}} \right] \\ &= \mathbb{E}_{\tau \sim \rho_{\theta}} \left[ \sum_{t=0}^{T-1} \left( \phi(s_t, a_t) - \mathbb{E}_{a' \sim \pi(s_t)} \phi(s_t, a') \right) A_t^{\pi_{\theta}}(s_t, a_t) \right] \\ &= \mathbb{E}_{\tau \sim \rho_{\theta}} \left[ \sum_{t=0}^{T-1} \phi(s_t, a_t) A_t^{\pi_{\theta}}(s_t, a_t) \right] \end{aligned}$$

Why can we drop the  $\mathbb{E} \phi(s_t, a')$  term? By linearity of expectation, consider the dropped term at a single timestep:  $\mathbb{E}_{\tau \sim \rho_{\theta}} [(\mathbb{E}_{a' \sim \pi(s_t)} \phi(s_t, a')) A_t^{\pi_{\theta}}(s_t, a_t)]$ . By Adam's Law, we can wrap the advantage term in a conditional expectation on the state  $s_t$ . Then we already know that  $\mathbb{E}_{a \sim \pi(s)} A_t^{\pi}(s, a) = 0$ , and so this entire term vanishes.

### 4.5.2 Neural policies

More generally, we could map states and actions to unnormalized scores via some parameterized function  $f_{\theta} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ , such as a neural network, and choose actions according to a softmax:

$$\pi_{\theta}^{\text{general}}(a|s) = \frac{\exp(f_{\theta}(s, a))}{\sum_{a'} \exp(f_{\theta}(s, a'))}.$$

The score can then be written as

$$\nabla \log \pi_{\theta}(a|s) = \nabla f_{\theta}(s, a) - \mathbb{E}_{a \sim \pi_{\theta}(s)} \nabla f_{\theta}(s, a')$$

### 4.5.3 Continuous action spaces

Consider a continuous  $n$ -dimensional action space  $\mathcal{A} = \mathbb{R}^n$ . Then for a stochastic policy, we could use a function to predict the *mean* action and then add some random noise about it. For example, we could use a neural network to predict the mean action  $\mu_{\theta}(s)$  and then add some noise  $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$  to it:

$$\pi_{\theta}(a|s) = \mathcal{N}(\mu_{\theta}(s), \sigma^2 I).$$

**Exercise:** Can you extend the “linear in features” policy to continuous action spaces in a similar way?

## 4.6 Local policy optimization

### 4.6.1 Motivation for policy gradient

Recall the policy iteration algorithm discussed in the MDP section : We alternate between these two steps:

- Estimating the  $Q$ -function of the current policy
- Updating the policy to be greedy w.r.t. this approximate  $Q$ -function.

(Note that we could equivalently estimate the advantage function.)

What advantages does the policy gradient algorithm have over policy iteration? Both policy gradient and policy iteration are iterative algorithms.

To analyze the difference between them, we’ll make use of the **performance difference lemma**.

#### Theorem 4.6.1: Performance difference lemma

Let  $\rho_{\pi, s}$  denote the distribution induced by the policy  $\pi$  over trajectories starting in state  $s$ .

Given two policies  $\pi, \tilde{\pi}$ , the PDL allows us to express the difference between their value functions as follows:

$$V_0^{\tilde{\pi}}(s) - V_0^{\pi}(s) = \mathbb{E}_{\tau \sim \rho_{\tilde{\pi}, s}} \left[ \sum_{h=0}^{H-1} A_h^{\pi}(s_h, a_h) \right] \quad (4.6)$$

Some intuition: Recall that  $A_h^{\pi}(s, a)$  tells us how much better the action  $a$  is in state  $s$  than average, supposing actions are chosen according to  $\pi$ . How much better is  $\tilde{\pi}$  than  $\pi$ ? To answer this, we break down the trajectory step-by-step. At each step, we compute how

cite

much better actions from  $\tilde{\pi}$  are than the actions from  $\pi$ . But this is exactly the average  $\pi$ -advantage, where the expectation is taken over actions from  $\tilde{\pi}$ . This is exactly what the PDL describes.

Let's analyze why fitted approaches such as PI don't work as well in the RL setting. To start, let's ask, where *do* fitted approaches work well? They are commonly seen in *supervised learning*, where a prediction rule is fit using some labelled training set, and then assessed on a test set from the same distribution. Does this assumption still hold when doing PI?

Let's consider a single iteration of PI. Suppose the new policy  $\tilde{\pi}$  chooses some action with a negative advantage w.r.t.  $\pi$ . Define  $\Delta_\infty = \min_{s \in \mathcal{S}} A_h^\pi(s, \tilde{\pi}(s))$ . If this is negative, then the PDL shows that there may exist some state  $s$  and time  $h$  such that

$$V_h^{\tilde{\pi}}(s) \geq V_h^\pi(s) - H \cdot |\Delta_\infty|.$$

In general, PI cannot avoid particularly bad situations where the new policy  $\tilde{\pi}$  often visits these bad states, causing an actual degradation. It does not enforce that the trajectory distributions  $\rho_\pi$  and  $\rho_{\tilde{\pi}}$  be close to each other. In other words, the “training distribution” that our prediction rule is fitted on,  $\rho_\pi$ , may differ significantly from the “evaluation distribution”  $\rho_{\tilde{\pi}}$  — we must address this issue of *distributional shift*.

How can we enforce that the *trajectory distributions* do not change much at each step? In fact, policy gradient already does this to a small extent: Supposing that the mapping from parameters to trajectory distributions is relatively smooth, then, by adjusting the parameters a small distance from the current iterate, we end up at a new policy with a similar trajectory distribution. But this is not very rigorous, and in practice the parameter-to-distribution mapping may not be smooth. Can we constrain the distance between the resulting distributions more explicitly? This brings us to the next two methods: **trust region policy optimization** (TRPO) and the **natural policy gradient** (NPG).

## 4.6.2 Trust region policy optimization

TRPO is another iterative algorithm for policy optimization. It is similar to policy iteration, except we constrain the updated policy to be “close to” the current policy in terms of the trajectory distributions they induce.

To formalize “close to”, we typically use the **Kullback-Leibler divergence (KLD)**:

### Definition 4.6.1: Kullback-Leibler divergence

For two PDFs  $p, q$ ,

$$\text{KL}(p \parallel q) := \mathbb{E}_{x \sim p} \left[ \log \frac{p(x)}{q(x)} \right] \quad (4.7)$$

This can be interpreted in many different ways, many stemming from information theory. Note that  $\text{KL}(p \parallel q) = 0$  if and only if  $p = q$ . Also note that it is not symmetric.



possible inter-  
actions

Additionally, rather than estimating the  $Q$ -function of the current policy, we can use the RHS of the Performance Difference Lemma (??) as our optimization target.

**Definition 4.6.2: Trust region policy optimization (exact)**

**Require:** Trust region radius  $\delta$

Initialize  $\theta^0$

**for**  $k = 0, \dots, K - 1$  **do**

$\theta^{k+1} \leftarrow \arg \max_{\theta} \mathbb{E}_{s_0, \dots, s_{H-1} \sim \pi^k} \left[ \sum_h \mathbb{E}_{a_h \sim \pi_{\theta}(s_h)} A^{\pi^k}(s_h, a_h) \right]$  ▷ See below

where  $\text{KL}(\rho_{\pi^k} \parallel \rho_{\pi_{\theta}}) \leq \delta$

**end for**

**return**  $\pi^K$

Note that the objective function is not identical to the r.h.s. of the Performance Difference Lemma. Here, we still use the *states* sampled from the old policy, and only use the *actions* from the new policy. This is because it would be computationally infeasible to sample entire trajectories from  $\pi_{\theta}$  as we are optimizing over  $\theta$ . This approximation is also reasonable in the sense that it matches the r.h.s. of the Performance Difference Lemma to first order in  $\theta$ . (We will elaborate more on this later.) See the TRPO paper for details.

add reference  
trpo

Both the objective function and the KLD constraint involve a weighted average over the space of all trajectories. This is intractable in general, so we need to estimate the expectation. As before, we can do this by taking an empirical average over samples from the trajectory distribution. However, the inner expectation over  $a_h \sim \pi_{\theta}$  involves the optimizing variable  $\theta$ , and we'd like an expression that has a closed form in terms of  $\theta$  to make optimization tractable. Otherwise, we'd need to resample many times each time we made an update to  $\theta$ . To address this, we'll use a common technique known as **importance sampling**.

**Definition 4.6.3: Importance sampling**

Suppose we want to estimate  $\mathbb{E}_{x \sim \tilde{p}}[f(x)]$ . However,  $\tilde{p}$  is difficult to sample from, so we can't take an empirical average directly. Instead, there is some other distribution  $p$  that is easier to sample from, e.g. we could draw samples from an existing dataset, as in the case of **offline RL**.

Then note that

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

so, given i.i.d. samples  $x_0, \dots, x_{N-1} \sim p$ , we can construct an unbiased estimate of  $\mathbb{E}_{x \sim \tilde{p}}[f(x)]$  by *reweighting* these samples according to the likelihood ratio  $\tilde{p}(x)/p(x)$ :

$$\frac{1}{N} \sum_{n=0}^{N-1} \frac{\tilde{p}(x_n)}{p(x_n)} f(x_n)$$

Doesn't this seem too good to be true? If there were no drawbacks, we could use this to estimate *any* expectation of any function on any arbitrary distribution! The drawback is that the variance may be very large due to the likelihood ratio term. If the sampling distribution  $p$  assigns low probability to any region where  $\tilde{p}$  assigns high probability, then the likelihood ratio will be very large and cause the variance to blow up.

Applying importance sampling allows us to estimate the TRPO objective as follows:

**Definition 4.6.4: Trust region policy optimization (implementation)**

Initialize  $\theta^0$

**for**  $k = 0, \dots, K - 1$  **do**

Sample  $N$  trajectories from  $\rho^k$  to learn a value estimator  $\tilde{b}_h(s) \approx V_h^{\pi^k}(s)$

Sample  $M$  trajectories  $\tau_0, \dots, \tau_{M-1} \sim \rho^k$

$$\theta^{k+1} \leftarrow \arg \max_{\theta} \frac{1}{M} \sum_{m=0}^{M-1} \sum_{h=0}^{H-1} \frac{\pi_{\theta}(a_h | s_h)}{\pi^k(a_h | s_h)} [R_h(\tau_m) - \tilde{b}_h(s_h)]$$

$$\text{where } \sum_{m=0}^{M-1} \sum_{h=0}^{H-1} \log \frac{\pi_k(a_h^m | s_h^m)}{\pi_{\theta}(a_h^m | s_h^m)} \leq \delta$$

**end for**

### 4.6.3 Natural policy gradient

Instead, we can solve an approximation to the TRPO optimization problem. This will link us back to the policy gradient from before. We take a first-order approximation to objective function and second-order approximation to the KLD constraint. (A full derivation is given in the appendix.)

This results in the optimization problem

$$\begin{aligned} & \max_{\theta} \nabla_{\theta} J(\pi_{\theta^k})^{\top} (\theta - \theta^k) \\ & \text{where } \frac{1}{2} (\theta - \theta^k)^{\top} F_{\theta^k} (\theta - \theta^k) \leq \delta \end{aligned} \tag{4.8}$$

where  $F_{\theta^k}$  is the **Fisher information matrix** defined below.

**Definition 4.6.5: Fisher information matrix**

Let  $p_{\theta}$  denote a parameterized distribution. Its Fisher information matrix  $F_{\theta}$  can be defined equivalently as:

$$\begin{aligned} F_{\theta} &= \mathbb{E}_{x \sim p_{\theta}} [(\nabla_{\theta} \log p_{\theta}(x)) (\nabla_{\theta} \log p_{\theta}(x))^{\top}] && \text{covariance matrix of the Fisher score} \\ &= \mathbb{E}_{x \sim p_{\theta}} [-\nabla_{\theta}^2 \log p_{\theta}(x)] && \text{average Hessian of the negative log-likelihood} \end{aligned}$$

Recall that the Hessian of a function describes its curvature: That is, for a vector  $\delta \in \Theta$ , the quantity  $\delta^\top F_\theta \delta$  describes how rapidly the negative log-likelihood changes if we move by  $\delta$ .

In particular, when  $p_\theta = \rho_\theta$  denotes a trajectory distribution, we can further simplify the expression:

$$F_\theta = \mathbb{E}_{\tau \sim \rho_\theta} \left[ \sum_{h=0}^{H-1} (\nabla \log \pi_\theta(a_h | s_h)) (\nabla \log \pi_\theta(a_h | s_h))^\top \right] \quad (4.9)$$

Note that we've used the Markov property to cancel out the cross terms corresponding to two different time steps.

This is a convex optimization problem, and so we can find the global optima by setting the gradient of the Lagrangian to zero:

$$\begin{aligned} \mathcal{L}(\theta, \eta) &= \nabla_\theta J(\pi_{\theta^k})^\top (\theta - \theta^k) - \eta \left[ \frac{1}{2} (\theta - \theta^k)^\top F_{\theta^k} (\theta - \theta^k) - \delta \right] \\ \nabla_\theta \mathcal{L}(\theta^{k+1}, \eta) &= 0 \\ \nabla_\theta J(\pi_{\theta^k}) &= \eta F_{\theta^k} (\theta^{k+1} - \theta^k) \\ \theta^{k+1} &= \theta^k + \eta F_{\theta^k}^{-1} \nabla_\theta J(\pi_{\theta^k}) \\ \text{where } \eta &= \sqrt{\frac{\delta}{\nabla_\theta J(\pi_{\theta^k})^\top F_{\theta^k} \nabla_\theta J(\pi_{\theta^k})}} \end{aligned}$$

include Lagrangian not in appendix

#### Definition 4.6.6: Natural policy gradient

**Require:** Learning rate  $\eta > 0$

Initialize  $\theta^0$

**for**  $k = 0, \dots, K - 1$  **do**

Estimate the policy gradient  $\hat{g} \approx \nabla_\theta J(\pi_{\theta^k})$

▷ See (??)

Estimate the Fisher information matrix  $\hat{F} \approx F_{\theta^k}$

▷ See (??)

$\theta^{k+1} \leftarrow \theta^k + \eta \hat{F}^{-1} \hat{g}$

▷ Natural gradient update

**end for**

How many trajectory samples do we need to accurately estimate the Fisher information matrix? As a rule of thumb, the sample complexity should scale with the dimension of the parameter space. This makes this approach intractable in the deep learning setting where we might have a very large number of parameters.

For some intuition: The typical gradient descent algorithm treats the parameter space as “flat”, treating the objective function as some black box value. However, in the case here where the parameters map to a *distribution*, using the natural gradient update is equivalent to optimizing over distribution space rather than parameter space.

**Example 4.6.1: Natural gradient on a simple problem**

Let's step away from reinforcement learning specifically and consider the following optimization problem over Bernoulli distributions  $\pi \in \Delta(\{0, 1\})$ :

$$J(\pi) = 100 \cdot \pi(1) + 1 \cdot \pi(0)$$

Clearly the optimal distribution is the constant one  $\pi(1) = 1$ . Suppose we optimize over the parameterized family  $\pi_\theta(1) = \frac{\exp(\theta)}{1 + \exp(\theta)}$ . Then our optimization algorithm should set  $\theta$  to be unboundedly large. Then the vanilla gradient is

$$\nabla_\theta J(\pi_\theta) = \frac{99 \exp(\theta)}{(1 + \exp(\theta))^2}.$$

Note that as  $\theta \rightarrow \infty$  that the increments get closer and closer to 0. However, if we compute the Fisher information scalar

$$\begin{aligned} F_\theta &= \mathbb{E}_{x \sim \pi_\theta} [(\nabla_\theta \log \pi_\theta(x))^2] \\ &= \frac{\exp(\theta)}{(1 + \exp(\theta))^2} \end{aligned}$$

resulting in the natural gradient update

$$\begin{aligned} \theta^{k+1} &= \theta^k + \eta F_{\theta^k}^{-1} \nabla_\theta J(\theta^k) \\ &= \theta^k + 99\eta \end{aligned}$$

which increases at a constant rate, i.e. improves the objective more quickly than vanilla gradient ascent.

### 4.6.4 Proximal policy optimization

Can we improve on the computational efficiency of the above methods?

We can relax the TRPO objective in a different way: Rather than imposing a hard constraint on the KL distance, we can instead impose a *soft* constraint by incorporating it into the objective:

**Definition 4.6.7: Proximal policy optimization (exact)**

**Require:** Regularization parameter  $\lambda$

Initialize  $\theta^0$

**for**  $k = 0, \dots, K - 1$  **do**

$$\theta^{k+1} \leftarrow \arg \max_{\theta} \mathbb{E}_{s_0, \dots, s_{H-1} \sim \pi^k} \left[ \sum_h \mathbb{E}_{a_h \sim \pi_\theta(s_h)} A^{\pi^k}(s_h, a_h) \right] - \lambda \text{KL}(\rho_{\pi^k} \parallel \rho_{\pi_\theta})$$

**end for**

**return**  $\theta^K$

Note that like the original TRPO algorithm ??, PPO is not gradient-based; rather, at each step, we try to maximize local advantage relative to the current policy.

Let us now turn this into an implementable algorithm, assuming we can sample trajectories from  $\pi_{\theta^k}$ .

Let us simplify the  $\text{KL}(\rho_{\pi^k} \parallel \rho_{\pi_\theta})$  term first. Expanding gives

$$\begin{aligned} \text{KL}(\rho_{\pi^k} \parallel \rho_{\pi_\theta}) &= \mathbb{E}_{\tau \sim \rho_{\pi^k}} \left[ \log \frac{\rho_{\pi^k}(\tau)}{\rho_{\pi_\theta}(\tau)} \right] \\ &= \mathbb{E}_{\tau \sim \rho_{\pi^k}} \left[ \sum_{h=0}^{H-1} \log \frac{\pi^k(a_h \mid s_h)}{\pi_\theta(a_h \mid s_h)} \right] && \text{state transitions cancel} \\ &= \mathbb{E}_{\tau \sim \rho_{\pi^k}} \left[ \sum_{h=0}^{H-1} \log \frac{1}{\pi_\theta(a_h \mid s_h)} \right] + c \end{aligned}$$

where  $c$  is some constant relative to  $\theta$ .

As we did for TRPO (??), we can use importance sampling (??) to rewrite the inner expectation. Combining the expectations together, this gives the (exact) objective

$$\max_{\theta} \mathbb{E}_{\tau \sim \rho_{\pi^k}} \left[ \sum_{h=0}^{H-1} \left( \frac{\pi_\theta(a_h \mid s_h)}{\pi^k(a_h \mid s_h)} A^{\pi^k}(s_h, a_h) - \lambda \log \frac{1}{\pi_\theta(a_h \mid s_h)} \right) \right]$$

Now we can use gradient ascent until convergence to maximize this function, completing a single iteration of PPO.

## 4.7 Guarantees for policy gradient methods

What kinds of problems are policy gradient methods good at solving?

For example, consider a very simple MDP in which