# Reinforcement Learning: Foundations

Shie Mannor and Yishay Mansour

September 2019

# Contents

# Chapter 1

# Introduction and Overview

## 1.1 Course Administration

### 1.1.1 classes

Lectures and recitation would be held virtually via zoom.

### 1.1.2 resources

The course web site is available through the `moodle`. The web site will include the lecture slides and information about recitations. Previous years web site are available at `http://rl-tau-2019.wikidot.com/` and `http://rl-tau-2018.wikidot.com/`.

### 1.1.3 Prerequisites

Introduction to Machine Learning (or an equivalent course).

### 1.1.4 requirements

The final grade would be composed from the following:

20% Homework (both theory and programming).

10% Final project: deep Reinforcement Learning on Atari game.

70% Final exam.

As usual, a student need to pass the final exam to pass the course.

## 1.2 Motivation for RL

In the recent years there is a renew interest in Reinforcement learning. The new interest is grounded in the emerging applications of Reinforcement learning.

Over the years reinforcement learning has proven to be highly successful for playing board games over a long history. Gerald Tesauro in 1992 developed the TD-gammon, which uses a two layer neural-network to achieve a high performance backgammon computer game. The network was trained by bootstrapping it from scratch and learned a temporal differences function. One of the amazing features of the TD-gammon was that even in the *first* move, it used a different game move than the grandmasters, who later adopted the new game move once they observe TD-gammon playing it.

Recently in 2015-7, Deep Mind have developed a deep neural-network to play Go, which is able to beat the best Go players in the world. Developing a human-level computer Go program has been a challenge that has persisted for decades.

Early in 1962, Artur Samuel developed a checkers game, which was at the level of the best human. His original framework included many of the ingredients which latter contributed to Reinforcement learning, as well as search heuristic for large domains.

To complete the picture of computer board games, we should mention Deep Blue, from 1996, which was able to beat the world champion Kasparov. This program was mainly built on a heuristic search, and developed a new simple hardware to support it. Recently, Deep Mind came out with a deep neural-network that matched the best chess programs (which are already much better than any human players).

Another domain, popularized by Deep Mind, is playing Atari video games, which where popular in the 1980's. Deep Mind were able to show how deep neural networks can achieve human level performance, using only the video picture as input (and having no additional information about the goal of the game).

Looking in to the future, the real promise of Reinforcement learning is learning in interactive settings. Probably one the most important applications is robotics, where reaching a human level performance would have far outreaching consequences.

*Origins of reinforcement learning:* Reinforcement learning spans a large number of disciplines. Naturally, we are going to look through the lens of Computer Science and Machine Learning. In Engineering, there are many disciplines related to optimal control. Other notable origins are in Operation Research, where the initial mathematical works have originated. Addition disciplines include: Neuroscience, Psychology and Economics.

*Mathematical Model* The main mathematical model we will use is Markov Decision

Process (MDP). The model tries to capture uncertainty in the dynamics of the environment, the actions and our knowledge. The main focus would be on decision making, namely, selecting actions. The evaluation would consider the long term effect of the actions, trading-off immediate rewards with long-term gains.

In contrast to Machine Learning, the reinforcement learning model would have a *state*, and the algorithm will influence the state through its actions. The algorithm would be faced with an inherent tradeoff between exploitation (getting the most reward given the current information) and exploration (gathering more information about the environment).

## 1.3 Markov Decision Process (MDP)

Our main formal model would be a Markov Decision Process (MDP) will be composed from:

- $\mathcal{S}$: a finite set of states.
- $\mathtt{s}_0 \in \mathcal{S}$: is the start state.
- $\mathcal{A}$: a finite set of actions.
- $\mathtt{p} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: a stochastic transition probability function, where $\Delta(\mathcal{S})$ is the set of probability distributions over $\mathcal{S}$. We denote by $\mathtt{p}(\cdot|\mathtt{s}, \mathtt{a})$ the distribution of next state, when doing action $\mathtt{a}$ in state $\mathtt{s}$.
- $\mathtt{R}$: an immediate reward. In general $\mathtt{R}$ would depend on the current state $\mathtt{s}$ and action $\mathtt{a}$, and in general it can be a random variable. In most cases we will focus on the expectation of $\mathtt{R}(\mathtt{s}, \mathtt{a})$. We will assume that the immediate reward $\mathtt{R}$ is bounded, specifically, that it is always in the range $[0, 1]$.

A run of the MDP is characterized by a *trajectory*, which has quadruples $(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}, \mathtt{r}_\mathtt{t}, \mathtt{s}_{\mathtt{t}+1})$, where $\mathtt{s}_\mathtt{t}$ is the state at time $\mathtt{t}$, $\mathtt{a}_\mathtt{t}$ is the action performed at time $\mathtt{t}$ in $\mathtt{s}_\mathtt{t}$, $\mathtt{r}_\mathtt{t}$ is the immediate reward, and $\mathtt{s}_{\mathtt{t}+1}$ is the next state. Clearly, $\mathtt{r}_\mathtt{t}$ is distributed according to $\mathtt{R}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})$ and $\mathtt{s}_{\mathtt{t}+1}$ is distributed according to $\mathtt{p}(\cdot|\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})$.

*Return function* We like to combine the immediate rewards to a single value, which we call *return*, that we will optimize. The decision to reduce all the immediate rewards to a single value is already a major modeling decision. When defining the return we will consider whether earlier immediate rewards are more important than later rewards. Also, there is an issue whether the system is *terminating*, namely terminates after a finite number of steps, or is *continuous*, which implies that it runs forever. Usually, the return would be linear in the immediate rewards, and we will be interested in maximizing expected return. For that reason, we will be mostly interested in the expectation of the immediate rewards.

Popular return functions include:

1. *Finite horizon:* There is a parameter $\mathtt{T}$ and the return is the sum of the first $\mathtt{T}$ rewards, i.e., $\sum_{\mathtt{t}=1}^{\mathtt{T}} \mathtt{r}_\mathtt{t}$.

2. *Infinite discounted return:* There is a parameter $\gamma \in (0, 1)$ and the discounted return is $\sum_{\mathtt{t}=0}^{\infty} \gamma^t \mathtt{r}_\mathtt{t}$. Note that since $\mathtt{r}_\mathtt{t} \in [0, 1]$, the return is bounded by $1/(1-\gamma)$.

3. *Average Reward:* where we take the limit of the average immediate reward. Specifically, $\lim_{\mathtt{T}\to\infty} \frac{1}{tHorizon} \mathbb{E}[\sum_{\mathtt{t}=1}^{\mathtt{T}} \mathtt{r}_\mathtt{t}]$.

4. *Sum of the rewards:* this applies only to the case of terminating MDP, since otherwise it might be infinite.

*Example of an MDP:* Consider an inventory control problem. At day $\mathtt{t}$ we have $x_\mathtt{t} \geq 0$ remaining items from previous days. We order $\mathtt{a}_\mathtt{t} \geq 0$ new items, which is our action. We have a demand of $d_\mathtt{t} \geq 0$, the amount consumer want to buy at day $\mathtt{t}$. We have $\mathtt{s}_\mathtt{t} = \min(d_\mathtt{t}, x_\mathtt{t} + \mathtt{a}_\mathtt{t})$ items bought. For day $\mathtt{t}+1$ we have $x_{\mathtt{t}+1}$ remaining items, i.e., $x_{\mathtt{t}+1} = \max(0, x_\mathtt{t} + \mathtt{a}_\mathtt{t} - d_\mathtt{t})$.

We can now formalize the immediate reward, based on $P$, the profit per item, $J(\cdot)$ the cost to order, and $C(\cdot)$ the cost of inventory. Our immediate reward is,

$$\mathtt{R}(x_\mathtt{t}, \mathtt{a}_\mathtt{t}) = P\mathtt{s}_\mathtt{t} - J(\mathtt{a}_\mathtt{t}) - C(x_{\mathtt{t}+1})$$

Our goal can be to maximize a discounted return function over the immediate rewards, where the discounting represents the interest rate.

*Action selection:* We assume that the state of the system is "observable", namely, we know in which state we are. Our goal would be to select action as to maximize the expected return. For the remainder of the lecture we focus on the discounted return.

Let a policy be a mapping from states to actions. Due to the Markov property of the system, we can show that the optimal history-dependent strategy is a deterministic policy, namely, it does not depend on the history and selects at each state a single action.

**Theorem 1.1.** *There exists a deterministic policy which maximizes the discounted return.*

Note that the optimal policy does not depend on the start state!

*Multi-Arm Bandits (MAB):* A simple well-studied variant of the MDP model are MABs, which are essentially an MDP with a single state. Given the model it is clear that the optimal policy would select the action with the highest immediate reward. However, when we need to learn the stochastic rewards, we have a tradeoff between using the action with the highest observed immediate reward versus trying new actions, and getting a better approximation of their expectation.

## 1.4 Planning

Planning problems capture the case that we are given a complete model of the MDP and would like to perform some task. Two popular tasks are the following:

- **Policy evaluation**: Given a policy $\pi$ evaluate its expected return.
- **Optimal control**: Compute an optimal policy $\pi^*$. For the infinite discounted return we have that $\pi^*$ maximizes the return from any start state.

We start with policy evaluation. An important ingredients in the planning process would be the following two value functions, which depend on the policy $\pi$.

- $\mathcal{V}^\pi(\mathtt{s})$, which is the expected return of $\pi$ starting from state $\mathtt{s}$.
- $\mathcal{Q}^\pi(\mathtt{s},\mathtt{a})$, which is the expected return of $\pi$ starting from state $\mathtt{s}$ doing action $\mathtt{a}$ and then following $\pi$.

We denote by $\mathcal{V}^*(\mathtt{s})$ and $\mathcal{Q}^*(\mathtt{s},\mathtt{a})$ the value function and the $\mathcal{Q}$-value function, respectively, of the optimal policy $\pi^*$. For the optimal policy we have,

$$\forall \mathtt{s} \in \mathcal{S} \quad \mathcal{V}^*(\mathtt{s}) = \max_{\pi} \mathcal{V}^\pi(\mathtt{s})$$

Namely, the optimal policy is optimal from any start state.

*Policy Evaluation* We can now solve the policy evaluation problem for the discounted return. We can write the following identities.

$$\forall \mathtt{s} \in \mathcal{S}: \quad \mathcal{V}^\pi(\mathtt{s}) = \mathbb{E}_{\mathtt{s}' \sim \mathtt{p}(s,\pi(\mathtt{s}))}[\mathtt{R}(s,\pi(\mathtt{s})) + \gamma \mathcal{V}^\pi(\mathtt{s}')]$$

This gives a system of linear equations where the unknowns are $\mathcal{V}^\pi(\mathtt{s})$. We have $|\mathcal{S}|$ equations and $|\mathcal{S}|$ unknowns, so there exists some solution.

*Optimal Control:* We can write the identity for the optimal $\mathcal{Q}$-function.

$$\mathcal{Q}^{\pi}(\mathtt{s},\mathtt{a}) = \mathbb{E}_{\mathtt{s}'\sim \mathtt{p}(s,\pi(\mathtt{s}))}[\mathtt{R}(\mathtt{s},\mathtt{a}) + \gamma \mathcal{V}^{\pi}(\mathtt{s}')]$$

Note that for a deterministic policy $\pi$ we have $\mathcal{V}^{\pi}(\mathtt{s}) = \mathcal{Q}^{\pi}(\mathtt{s},\pi(\mathtt{s}))$.

The following theorem gives a characterization of the optimal policy (also known as Bellman Eq).

**Theorem 1.2.** *A policy $\pi$ is optimal if and only if at each state $\mathtt{s}$ we have*

$$\mathcal{V}^{\pi}(\mathtt{s}) = \max_{\mathtt{a}}\{\mathcal{Q}^{\pi}(\mathtt{s},\mathtt{a})\}$$

*Proof.* We will show here only the *only if* part (the other direction will be give later in the course). Assume that there is a state $\mathtt{s}$ and action $\mathtt{a}$ such that

$$\mathcal{V}^{\pi}(\mathtt{s}) < \mathcal{Q}^{\pi}(\mathtt{s},\mathtt{a})$$

The strategy of performing in state $\mathtt{s}$ action $\mathtt{a}$ and then using policy $\pi$ outperforms policy $\pi$, and so policy $\pi$ is not optimal.

The improvement would be valid for each visit of state $\mathtt{s}$, so the policy that performs action $\mathtt{a}$ in state $\mathtt{s}$ improves over policy $\pi$ (and is also a policy, a mapping from states to actions). □

*Computing the optimal policy:* There are three popular algorithms to compute the optimal policy given an MDP model.

- *Linear Program*
- *Value Iteration*: at iteration $\mathtt{t}$ compute

  $$V_{\mathtt{t}+1}(\mathtt{s}) = \max_{\mathtt{a}} \mathbb{E}_{\mathtt{s}'\sim \mathtt{p}(\cdot|\mathtt{a},\mathtt{s})}[\mathtt{R}(\mathtt{s},\pi(\mathtt{s})) + \gamma V_{\mathtt{t}}(\mathtt{s}')]$$

  We will show that at each iteration the distance from the optimal value function $\mathcal{V}^*$ is decreased by $1-\gamma$, namely $\|V_{\mathtt{t}+1} - \mathcal{V}^*\|_\infty \le (1-\gamma)\|V_{\mathtt{t}} - \mathcal{V}^*\|_\infty$.
- *Policy Iteration*:

  $$\pi_{\mathtt{t}+1}(\mathtt{s}) = \arg\max_{\mathtt{a}}\{\mathcal{Q}^{\pi_{\mathtt{t}}}(\mathtt{s},\mathtt{a})\}.$$

  We will show that the number of iterations of policy iteration is bounded by that of value iteration, but each iteration is computationally more intensive.

## 1.5 Learning Algorithms

We now would like to address the case that the MDP model is unknown. We still have the two primary tasks: (1) policy evaluation, and (2) optimal control.

We will use two different approaches. The *model based* approach, first learns a model and then uses it. The *model free* approach learns directly a policy.

### 1.5.1 Model Based Learning

The basic idea is very intuitive. We like to estimate the model from observations. Given a trajectory, we can decompose it to quadruplets $(s_t, a_t, r_t, s_{t+1})$. We can learn both the immediate rewards $R(s, a)$ and the transition function $p(\cdot|s, a)$ from those quadruplets.

*Building the observed model (off-policy):* Given $(s_t, a_t, r_t, s_{t+1})$ we define the observed model as follows. Let $\#(s, a) = \sum_{t=1}^{T} \mathbb{I}(s_t = s, a_t = a)$, where $\mathbb{I}(\cdot)$ is the indicator function. The observed reward would be

$$\widehat{R}(s, a) = \sum_{t=1}^{T} r_t \frac{\mathbb{I}(s_t = s, a_t = a)}{\#(s, a)}.$$

The observed next state distribution would be:

$$\widehat{p}(s'|s, a) = \frac{\sum_{t=1}^{T} \mathbb{I}(s_t = s, a_t = a, s_{t+1} = s')}{\#(s, a)}.$$

Given an observed model, we can compute an optimal policy for the observed model. The following intuitive claim can be (and would be) made formal (later in the course).

**Claim 1.3.** *If the observed model is "accurate" then the optimal policy for the observed model is a near optimal policy for the true model.*

There is a hidden assumption that we have enough samples for each state $s$ and action $a$. An important question is how many samples we need for each $(s, a)$ to get an accurate observed model.

*Building the observed model (on-policy):* In an on-policy the learner can control the actions. How can it use this ability to accelerate the learning. The simple idea is to try to visit states which we have not visit sufficiently, which will help us to build an accurate observed model.

A basic idea is to split the states to two parts. Well observed states, from which we have sufficient samples for each action. Relatively unknown states, from which we have not sampled sufficiently. The idea is that from the well sampled states, we have a good model. Now we can "imagine" that the immediate reward in the relatively unknown states is maximal, while the immediate rewards in the well observed states is zero. We will also assume that once we get to a relatively unknown state we stay in such a state. Given such a model, we can solve for the planning problem. The optimal policy for this (imaginary) model will find a shortest path to the relatively unknown states, giving us an additional observation for those states. Eventually, states would move from the relatively unknown states to the well known states. Once the set of relatively unknown states is empty, we are done with the learning phase.

*Monte-Carlo Methods* For those methods it is easiest to think of a terminating MDP which works in episodes. The Monte Carlo algorithm runs an episode using the policy. Given the trajectories we like to build an observed model. However, there might be statistical issues. Unlike the continuous trajectory, now only the first arrival to a state is an independent sample! Additional visits to the state might be correlated with previous outcomes!

### 1.5.2 Model free learning

Model free learning algorithm try to learn directly the value function or the policy, and circumvent learning the model.

There is a variety of model free learning algorithms. They differ by the value function they estimate, $\mathcal{Q}$ or $\mathcal{V}$, and whether they are off-policy or on-policy. The main challenge is to analyze their dynamics and guarantee their convergence.

### 1.5.3 $\mathcal{Q}$-learning: off-policy

The $\mathcal{Q}$-learning algorithm estimates directly the optimal learning function. It receives as input a long trajectory and outputs an estimate for the $\mathcal{Q}$ function.

The idea is to focus on the difference between the current estimate and the observed values. Given the time $\mathtt{t}$ quadruple $(\mathtt{s_t}, \mathtt{a_t}, \mathtt{r_t}, \mathtt{s_{t+1}})$, we define

$$\Delta_{\mathtt{t}} = Q_{\mathtt{t}}(\mathtt{s_t}, \mathtt{a_t}) - \mathtt{r_t} - \gamma \max_{\mathtt{a}} Q_{\mathtt{t}}(\mathtt{s_{t+1}}, \mathtt{a})$$

We have a learning rate $\alpha_{\mathtt{t}}(\mathtt{s}, \mathtt{a})$ which may depend on the number of times, until time $\mathtt{t}$, we executed $(\mathtt{s}, \mathtt{a})$. We now update our estimate to $\mathcal{Q}^{\mathtt{t}+1}$ as follows,

$$Q_{\mathtt{t}+1}(\mathtt{s_t}, \mathtt{a_t}) = Q_{\mathtt{t}+1}(\mathtt{s_t}, \mathtt{a_t}) - \alpha(\mathtt{s_t}, \mathtt{a_t})\Delta_{\mathtt{t}}$$

Note that once $Q_{\mathtt{t}} = \mathcal{Q}^*$ then $E[\Delta_{\mathtt{t}}] = 0$. The challenge in the analysis is to study the dynamics of the stochastic process and show the convergence.

### 1.5.4 Temporal Differences

The Temporal Differences (TD) computes an estimate to the $V$ value function of the current policy. The error at time $\mathtt{t}$ is

$$\Delta_{\mathtt{t}} = V_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}) - \mathtt{r}_{\mathtt{t}} - V_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}+1})$$

and, using the learning rate $\alpha_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}})$ we update

$$V_{\mathtt{t}+1}(\mathtt{s}_{\mathtt{t}}) = V_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}) - \alpha_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}})\Delta_{\mathtt{t}}$$

The above is called $TD(0)$, which focuses on the last transition. In general we can add a parameter $\lambda$ which defines $TD(\lambda)$. The parameter allows us to update the current value also using recent observed rewards. The $\lambda$ can be viewed as a discounting, which is used for the update (and not the return!). The eligibility of a state counts how many times a state is visited (discounted by $\lambda$). The eligibility trace of a state $\mathtt{s}$ is

$$e_{\mathtt{t}}(\mathtt{s}) = \sum_{i=1}^{\mathtt{t}} (\lambda\gamma)^{\mathtt{t}-i}\mathbb{I}(\mathtt{s}_{\mathtt{t}-i} = \mathtt{s}) = (\lambda\gamma)e_{\mathtt{t}-1}(\mathtt{s}) + \mathbb{I}(\mathtt{s}_{\mathtt{t}} = \mathtt{s})$$

Given the eligibility trace the new estimated value, in *every* state $\mathtt{s}$, is

$$V_{\mathtt{t}+1}(\mathtt{s}) = V_{\mathtt{t}}(\mathtt{s}) - \alpha_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}})\Delta_{\mathtt{t}} e_{\mathtt{t}}(\mathtt{s})$$

The idea is that we propagate the rewards faster to recently visited states.

## 1.6 Large state MDP

In the previous learning algorithms we assumed that the number of states is sufficiently small, since we build lookup table index by states. However, in many applications the number of states is exponential in the number of natural parameters. Overcoming this challenge can be done in many ways, here are a few popular ones.

1. *Restricted value function:* The main idea is to use a value function from a limited class. Two extreme solutions are linear functions and deep neural networks. This is similar to supervised learning, where we learn using a given function class. Especially popular are Deep Q-Networks (DQN) which learn the $\mathcal{Q}$ values using a deep neural network.

2. *Restricted policy class:* We fix a policy class $\Pi = \{\pi : \mathcal{S} \to \mathcal{A}\}$. The main challenge is given $\pi \in \Pi$ to estimate $\mathcal{V}^\pi$ or $\mathcal{Q}^\pi$. Given the estimate of $\mathcal{Q}^\pi$ we can improve the policy by computing a greedy policy $\pi' = \arg\max \mathcal{Q}^\pi$. The quality clearly depends on the approximation of both the $\mathcal{Q}^\pi$ and the ability to fit $\pi' = \arg\max \mathcal{Q}^\pi$ to $\Pi$.

   One challenge is how to compute the gradient of a policy, to update its parameters. The challenge is that the update influences not only the action probabilities, but also the distribution of the states.

3. *Restricted MDP model:* We can make assumptions about the MDP structure, for example, MAB assumes a single state.

4. *Generative model:* We are given an implicit representation of the MDP using a generative model. The model, given $(\mathtt{s}, \mathtt{a})$ return $(\mathtt{r}, \mathtt{s}')$, appropriately distributed.

## 1.7 Course Schedule

- Part 1: MDP basics and planning
  - Deterministic Decision Processes: Finite Horizon and Average cost
  - Markov Chains and Markov Decision Processes: Finite Horizon
  - MDP: discounted infinite horizon
- Part 2: MDP learning Model Based and Model free
  - Model based learning: off-policy, on-policy, Rmax
  - Model free learning: Q-learning, SARSA, Monte-Carlo
  - Model free learning: TD(0), TD($\lambda$), Importance Sampling, Action-Critic
- Part 3: Large state MDP Policy Gradient, Deep Q-Network
  - Value Function Approximation
  - Policy Gradient
- Part 4: Special MDPs
  - Stochastic Multi-Arm Bandits

  - Partially Observable MDP (POMDP)
  - Linear Dynamics models (LQR)
  - Generative model; Inverse RL

# Chapter 2

# Deterministic Decision Processes

In this chapter we introduce the dynamic system viewpoint of the optimal planning problem. We restrict the discussion here to deterministic (rather than stochastic) systems. We consider two basic settings. The finite-horizon decision problem and its recursive solution via finite-horizon Dynamic Programming. The average cost and it related minimum average weight cycle.

## 2.1 Discrete Dynamic Systems

We consider a discrete-time dynamic system, of the form:

$$\mathtt{s}_{\mathtt{t}+1} = f_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}), \quad \mathtt{t} = 0, 1, 2, \ldots, \mathtt{T} - 1$$

where

- $\mathtt{t}$ is the time index.
- $\mathtt{s}_{\mathtt{t}} \in \mathcal{S}_{\mathtt{t}}$ is the state variable at time $\mathtt{t}$, and $\mathcal{S}_{\mathtt{t}}$ is the set of possible states at time $\mathtt{t}$.
- $\mathtt{a}_{\mathtt{t}} \in \mathcal{A}_{\mathtt{t}}$ is the control variable at time $\mathtt{t}$, and $\mathcal{A}_{\mathtt{t}}$ is the set of possible control actions at time $\mathtt{t}$.
- $f_{\mathtt{t}} : \mathcal{S}_{\mathtt{t}} \times \mathcal{A}_{\mathtt{t}} \to \mathcal{S}_{\mathtt{t}+1}$ is the state transition function, which defines the *state dynamics* at time $\mathtt{t}$.
- $\mathtt{T} > 0$ is the *time horizon* of the system. It can be finite or infinite.

**Remark 2.1.** *More generally, the set* $\mathcal{A}_{\mathtt{t}}$ *of available actions may depend on the state at time* $\mathtt{t}$*, namely:* $\mathtt{a}_{\mathtt{t}} \in \mathcal{A}_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}) \subset \mathcal{A}_{\mathtt{t}}$.

**Remark 2.2.** *The system is, in general, time-varying. It is called* time invariant *if* $f_{\mathtt{t}}, \mathcal{S}_{\mathtt{t}}, \mathcal{A}_{\mathtt{t}}$ *do not depend on the time* $\mathtt{t}$*. In that case we write*

$$\mathtt{s}_{\mathtt{t}+1} = f(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}), \quad \mathtt{t} = 0, 1, 2, \ldots, \mathtt{T}-1; \quad \mathtt{s}_{\mathtt{t}} \in \mathcal{S},\ \mathtt{a}_{\mathtt{t}} \in \mathcal{A}(\mathtt{s}_{\mathtt{t}}).$$

**Remark 2.3.** *The state dynamics may be augmented by an output equation:*

$$\mathtt{o}_{\mathtt{t}} = \mathcal{O}_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}),$$

*where* $\mathtt{o}_{\mathtt{t}}$ *is the system observation, or the output. In most of this book we implicitly assume that* $\mathtt{o}_{\mathtt{t}} = \mathtt{s}_{\mathtt{t}}$*, namely, the current state* $\mathtt{s}_{\mathtt{t}}$ *is fully observed.*

**Example 2.1.** ***Linear Dynamic Systems***

*A well known example of a dynamic system is that of a linear time-invariant system, where:*

$$\mathtt{s}_{\mathtt{t}+1} = A\mathtt{s}_{\mathtt{t}} + B\mathtt{a}_{\mathtt{t}}$$

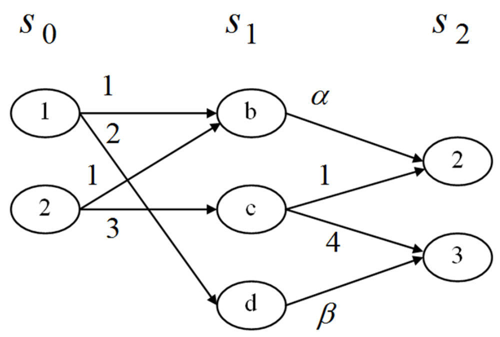
*with* $\mathtt{s}_{\mathtt{t}} \in \mathbb{R}^n$, $\mathtt{a}_{\mathtt{t}} \in \mathbb{R}^m$, $A \in \mathbb{R}^{n \times n}$ *and* $B \in \mathbb{R}^{n \times m}$*. Here the state and action spaces are evidently continuous (and not discrete).*

**Example 2.2.** ***Finite models***

*Our emphasis here will be on* finite state and action *models. A finite state space contains a finite number of points:* $\mathcal{S}_{\mathtt{t}} = \{1, 2, \ldots, n_{\mathtt{t}}\}$*. Similarly, a finite action space implies a finite number of control actions at each stage:*

$$\mathcal{A}_{\mathtt{t}}(\mathtt{s}) = \{1, 2, \ldots, m_{\mathtt{t}}(\mathtt{s})\}, \ \ \mathtt{s} \in \mathcal{S}_{\mathtt{t}}$$

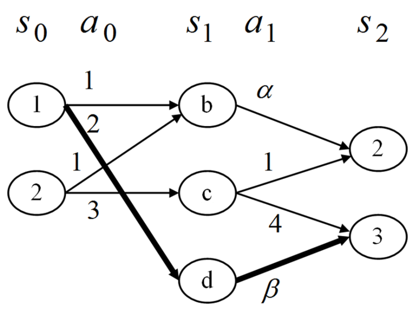
**Graphical description:** Finite models (over finite time horizons) can be represented by a corresponding decision graph:



Here:

- $\mathtt{T} = 2$, $\mathcal{S}_0 = \{1,2\}$, $\mathcal{S}_1 = \{b, c.d\}$, $\mathcal{S}_2 = \{2,3\}$,
- $\mathcal{A}_0(1) = \{1,2\}$, $\mathcal{A}_0(2) = \{1,3\}$, $\mathcal{A}_1(b) = \{\alpha\}$, $\mathcal{A}_1(c) = \{1,4\}$, $\mathcal{A}_1(d) = \{\beta\}$
- $f_0(1,1) = b$, $f_0(1,2) = d$, $f_0(2,1) = b$, $f_0(2,3) = c$, $f_1(b,\alpha) = 2$, etc.

**Definition 2.1.** ***Feasible Path***
*A feasible path for the specified system is a sequence* $(\mathtt{s}_0, \mathtt{a}_0, \ldots, \mathtt{s}_{\mathtt{T}-1}, \mathtt{a}_{\mathtt{T}-1}, \mathtt{s}_\mathtt{T})$ *of states and actions, such that* $\mathtt{s}_\mathtt{t} \in \mathcal{A}_\mathtt{t}(\mathtt{s}_\mathtt{t})$ *and* $\mathtt{s}_{\mathtt{t}+1} = f_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})$.



A feasible path

## 2.2 The Finite Horizon Decision Problem

We proceed to define our first and simplest planning problem. For that we need to specify a *performance objective* for our model, and the notion of *control policies*.

### 2.2.1 Costs and Rewards

**The cumulative cost:** Let $\mathtt{h}_\mathtt{T} = (\mathtt{s}_0, \mathtt{a}_0, \ldots, \mathtt{a}_{\mathtt{T}-1}, \mathtt{s}_{\mathtt{T}-1}, \mathtt{s}_\mathtt{T})$ denote an T-stage feasible path for the system. Each feasible path $\mathtt{h}_\mathtt{T}$ is assign some cost $\mathcal{C}_\mathtt{T} = \mathcal{C}_\mathtt{T}(\mathtt{h}_\mathtt{T})$.

The standard definition of the cost $\mathcal{C}_\mathtt{T}$ is through the following *cumulative cost functional*:

$$\mathcal{C}_\mathtt{T}(\mathtt{h}_\mathtt{T}) = \sum_{\mathtt{t}=0}^{\mathtt{T}-1} \mathtt{c}_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) + c_\mathtt{T}(\mathtt{s}_\mathtt{T})$$

Here:

- $\mathtt{c}_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})$ is the *instantaneous* cost or *single-stage* cost at stage t, and $\mathtt{c}_\mathtt{t}$ is the instantaneous cost function.
- $\mathtt{c}_\mathtt{T}(\mathtt{s}_\mathtt{T})$ is the *terminal* cost, and $\mathtt{c}_\mathtt{T}$ is the terminal cost function.

**Note:**

- The cost functional defined above is *additive* in time. Other cost functionals are possible, for example the max cost, but additive cost is by far the most common and useful.
- We shall refer to $\mathcal{C}_{\mathtt{T}}$ as the *cumulative* $\mathtt{T}$*-stage cost*, or just the *cumulative cost.*

Our objective is to *minimize* the cumulative cost $\mathcal{C}_{\mathtt{T}}$, by a proper choice of actions. We will define that goal more formally in the next section.

**Cost versus reward formulation:** It is often more natural to consider *maximizing* reward rather than minimizing cost. In that case, we define the cumulative $\mathtt{T}$-stage return function:

$$\mathcal{V}_{\mathtt{T}}(h_{\mathtt{T}}) = \sum_{\mathtt{t}=0}^{\mathtt{T}-1} \mathtt{r}_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) + \mathtt{r}_{\mathtt{T}}(\mathtt{s}_{\mathtt{T}})$$

Here and $\mathtt{r}_{\mathtt{t}}$ is the instantaneous reward, and $\mathtt{r}_{\mathtt{T}}$ is the terminal reward. Clearly, minimizing $\mathcal{C}_{\mathtt{T}}$ is equivalent to maximizing $\mathcal{V}_{\mathtt{T}}$, if we set:

$$\mathtt{r}_{\mathtt{t}}(\mathtt{s}, \mathtt{a}) = -\mathtt{c}_{\mathtt{t}}(\mathtt{s}, \mathtt{a}) \text{ and } \mathtt{r}_{\mathtt{T}}(\mathtt{s}) = -\mathtt{c}_{\mathtt{T}}(\mathtt{s}).$$

We denote by $\mathbb{T}$ the set of time steps input setting, i.e., $\mathbb{T} = \{1, \dots, \mathtt{T}\}$

### 2.2.2 Optimal Paths

Our first planning problem is the following $\mathtt{T}$*-stage Finite Horizon Problem*:

- For a given initial state $\mathtt{s}_0$, find a feasible path $\mathtt{h}_{\mathtt{T}} = (\mathtt{s}_0, \mathtt{a}_0, \dots, \mathtt{s}_{\mathtt{T}-1}, \mathtt{a}_{\mathtt{T}-1}, \mathtt{s}_{\mathtt{T}})$ that minimizes the cost functional $\mathcal{C}_{\mathtt{T}}(\mathtt{h}_{\mathtt{T}})$, over all feasible paths $\mathtt{h}_{\mathtt{T}}$.

Such a feasible path $\mathtt{h}_{\mathtt{T}}$ is called an *optimal path* from $\mathtt{s}_0$.

A more general notion than a path is that of a *control policy*, that specifies the action to be taken at each state. Control policies will play an important role in our Dynamic Programming algorithms, and are defined next.

### 2.2.3 Control Policies

In general we will consider a few classes of control policies. The two basic dimensions in which we will characterize the control policies is their dependence on the history, and their use of randomization.

- A general or **history-dependent** control policy $\pi = (\pi_{\mathtt{t}})_{\mathtt{t}\in\mathbb{T}}$ is a mapping from each possible history $\mathtt{h_t} = (\mathtt{s}_0, \mathtt{a}_0, \ldots, \mathtt{s}_{\mathtt{t}-1}, \mathtt{a}_{\mathtt{t}-1}, \mathtt{s_t})$, $\mathtt{t} \in \mathbb{T}$, to an action $\mathtt{a_t} = \pi_{\mathtt{t}}(\mathtt{h_t}) \in \mathcal{A}_{\mathtt{t}}$. We denote the set of general policies by $\Pi_H$.

- A **Markov** control policy $\pi$ is allowed to depend on the current state and time only: $\mathtt{a_t} = \pi_{\mathtt{t}}(\mathtt{s_t})$. We denote the set of Markov policies by $\Pi_M$.

- For stationary models, we may define **stationary** control policies that depend on the current state alone. A stationary policy is defined by a single mapping $\pi : \mathcal{S} \to \mathcal{A}$, so that $\mathtt{a_t} = \pi(\mathtt{s_t})$ for all $\mathtt{t} \in \mathbb{T}$. We denote the set of stationary policies by $\Pi_S$.

- Evidently, $\Pi_H \supset \Pi_M \supset \Pi_S$.

**Randomized (Stochastic) Control policies**

- The control policies defined above specify deterministically the action to be taken at each stage. In some cases we want to allow for a random choice of action.

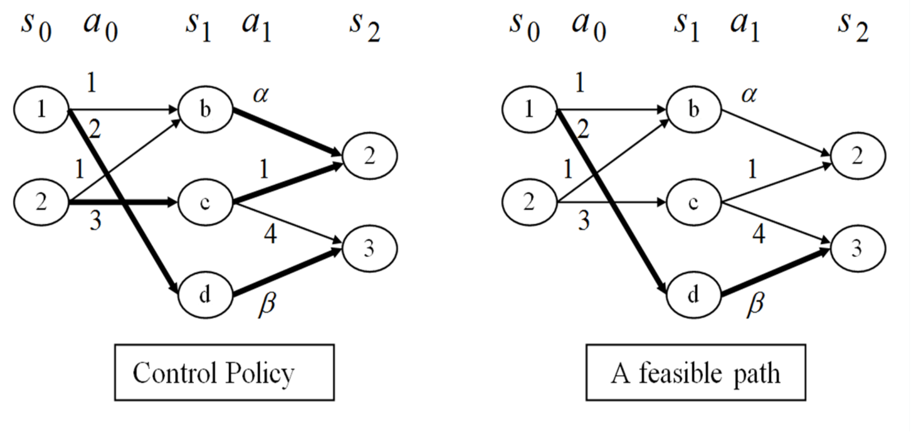
- A general randomized (stochastic) control policy assigns to each possible history $\mathtt{h_t}$ a probability distribution $\pi_{\mathtt{t}}(\cdot|\mathtt{h_t})$ over the action set $\mathcal{A}_{\mathtt{t}}$. That is, $\Pr\{\mathtt{a_t} = \mathtt{a}|\mathtt{h_t}\} = \pi_{\mathtt{t}}(\mathtt{a}|\mathtt{h_t})$. We denote the set of general randomized policies by $\Pi_{HS}$.

- Similarly, we can define the set $\Pi_{MS}$ of Markov randomized (stochastic) control policies, where $\pi_{\mathtt{t}}(\cdot|\mathtt{h_t})$ is replaced by $\pi_{\mathtt{t}}(\cdot|\mathtt{s_t})$, and the set $\Pi_{SS}$ of stationary randomized (stochastic) control policies, where $\pi_{\mathtt{t}}(\cdot|\mathtt{s_t})$ is replaced by $\pi(\cdot|\mathtt{s_t})$.

- Note that the set $\Pi_{HS}$ includes all other policy sets as special cases.

- For deterministic control policies, we similarly define $\Pi_{HD} \supset \Pi_{MD} \supset \Pi_{SD}$

**Control policies and paths:** As mentioned, a control policy specifies an action for each state, whereas a path specifies an action only for states along the path. The definition of a policy, allows us to consider counter-factual events, namely, what would have been the path if we considered a different action. This distinction is illustrated in the following figure.



**Induced Path:** A control policy $\pi$, together with an initial state $s_0$, specify a feasible path $h_T = (s_0, a_0, \ldots, s_{N-1}, a_{T-1}, s_T)$. This path may be computed recursively using $a_t = \pi_t(s_t)$ and $s_{t+1} = f_t(s_t, a_t)$, for $t = 0, 1, \ldots, T-1$.

**Remark 2.4.** *Suppose that for each state* $s_t$*, each action* $a_t \in \mathcal{A}_t(s_t)$ *leads to a different state* $s_{t+1}$ *(i.e., at most one edge connects any two states). We can then identify each action* $a_t \in \mathcal{A}_t(s_t)$ *with the next state* $s_{t+1} = f_t(s_t, a_t)$ *it induces. In that case a path may be uniquely specified by the state sequence* $(s_0, s_1, \ldots, s_T)$.

### 2.2.4 Reduction between control policies classes

We first show a reduction from a general history dependent policies to Randomized Markovian policies. The main observation is that the only influence on the cumulative cost is the expected instantaneous cost $\mathbb{E}[c_t(s_t, a_t)]$. Namely, let

$$\rho_t^\pi(s, a) = \Pr_{h'_{t-1}}[a_t = a, s_t = s] = \mathbb{E}_{h'_{t-1}}[\mathbb{I}[s_t = s, a_t = a]|h'_{t-1}],$$

where $h'_{t-1} = (s_0, a_0, \ldots, s_{t-1}, a_{t-1})$ is the history of the first $t-1$ time steps generated using $\pi$, and the probability and expectation are taken with respect to the randomness of the policy $\pi$. Now we can rewrite the expected cost to go as,

$$\mathbb{E}[\mathcal{C}^{\pi}(\mathtt{s}_0)] = \mathbb{E}[\sum_{\mathtt{t}=1}^{\mathtt{T}-1} \sum_{\mathtt{a}\in\mathcal{A}_\mathtt{t},\mathtt{s}\in\mathcal{S}_\mathtt{t}} \mathtt{c}_\mathtt{t}(\mathtt{s},\mathtt{a})\rho^{\pi}_\mathtt{t}(\mathtt{s},\mathtt{a})],$$

where $\mathcal{C}^{\pi}(\mathtt{s}_0)$ is the random variable of the cost when starting at state $\mathtt{s}_0$ and following policy $\pi$.

This implies that any two policies $\pi$ and $\pi'$ for which $\rho^{\pi}_\mathtt{t}(\mathtt{s},\mathtt{a}) = \rho^{\pi'}_\mathtt{t}(\mathtt{s},\mathtt{a})$, for any time $\mathtt{t}$, state $\mathtt{s}$ and action $\mathtt{a}$, would have the same expected cumulative cost for any cost function, i.e., $\mathbb{E}[\mathcal{C}^{\pi}(\mathtt{s}_0)] = \mathbb{E}[\mathcal{C}^{\pi'}(\mathtt{s}_0)]$

**Theorem 2.1.** *For any policy $\pi \in \Pi_{HS}$, there is a policy $\pi' \in \Pi_{MS}$, such that for every state $\mathtt{s}$ and action $\mathtt{a}$ we have, $\rho^{\pi}(\mathtt{s},\mathtt{a}) = \rho^{\pi'}(\mathtt{s},\mathtt{a})$. This will imply that,*

$$\mathbb{E}[\mathcal{C}^{\pi}(\mathtt{s}_0)] = \mathbb{E}[\mathcal{C}^{\pi'}(\mathtt{s}_0)]$$

*Proof.* Given the policy $\pi \in \Pi_{HS}$, we define $\pi' \in \Pi_{MS}$ as follows. For every state $\mathtt{s} \in \mathcal{S}_\mathtt{t}$ we define

$$\pi'_\mathtt{t}(\mathtt{a}|\mathtt{s}) = \Pr_{\mathtt{h}_{\mathtt{t}-1}}[\mathtt{a}_\mathtt{t} = \mathtt{a}|\mathtt{s}_\mathtt{t} = \mathtt{s}] = \frac{\rho^{\pi}_\mathtt{t}(\mathtt{s},\mathtt{a})}{\sum_{\mathtt{a}'\in\mathcal{A}_\mathtt{t}} \rho^{\pi}_\mathtt{t}(\mathtt{s},\mathtt{a}')}$$

By definition $\pi'$ is Markovian (depends only on the time $\mathtt{t}$ and the realized state $\mathtt{s}$). By construction $\rho^{\pi}_\mathtt{t}(\mathtt{s},\mathtt{a}) = \rho^{\pi'}_\mathtt{t}(\mathtt{s},\mathtt{a})$ is identical, which implies that $\mathbb{E}[\mathcal{C}^{\pi}(\mathtt{s}_0)] = \mathbb{E}[\mathcal{C}^{\pi'}(\mathtt{s}_0)]$. □

Next we show that for any stochastic Markovian policy there is a deterministic Markovian policy with at most the same cumulative cost.

**Theorem 2.2.** *For any policy $\pi \in \Pi_{MS}$, there is a policy $\pi' \in \Pi_{MD}$, such that*

$$\mathbb{E}[\mathcal{C}^{\pi}(\mathtt{s}_0)] \geq \mathbb{E}[\mathcal{C}^{\pi'}(\mathtt{s}_0)]$$

*Proof.* The proof is by backward induction on the steps. The inductive claim is:
*For any policy $\pi \in \Pi_{MS}$ which is deterministic in $[\mathtt{t}+1,\mathtt{T}]$, there is a policy $\pi' \in \Pi_{MS}$ which is deterministic in $[\mathtt{t},\mathtt{T}]$ and $\mathbb{E}[\mathcal{C}^{\pi}(\mathtt{s}_0)] \geq \mathbb{E}[\mathcal{C}^{\pi'}(\mathtt{s}_0)]$.*
Clearly, the theorem follows from the case of $\mathtt{t} = 0$.

For the base of the induction we can take $\mathtt{t} = \mathtt{T}$, which holds trivially.

For the inductive step, assume that $\pi \in \Pi_{MS}$ is deterministic in $[\mathtt{t}+1,\mathtt{T}]$.

For every $\mathtt{s}_{\mathtt{t}+1} \in \mathcal{S}_{\mathtt{t}+1}$ define

$$\mathcal{C}_{\mathtt{t}+1}(\mathtt{s}_{\mathtt{t}+1}) = \mathcal{C}(path(\mathtt{s}_{\mathtt{t}+1},\ldots,\mathtt{s}_\mathtt{T})),$$

where $path(\mathtt{s}_{\mathtt{t}+1},\ldots,\mathtt{s}_{\mathtt{T}})$ is the deterministic path from $\mathtt{s}_{\mathtt{t}+1}$ induced by $\pi$.

We define $\pi'$ to be identical to $\pi$ for all time steps $\mathtt{t}' \neq \mathtt{t}$. We define $\pi'_{\mathtt{t}}$ for each $\mathtt{s}_{\mathtt{t}} \in \mathcal{S}_{\mathtt{t}}$ as follows:

$$\pi'_{\mathtt{t}}(\mathtt{a}_{\mathtt{t}},\mathtt{s}_{\mathtt{t}}) = \arg\min_{\mathtt{a}\in\mathcal{A}_{\mathtt{t}}} \mathtt{c}(\mathtt{s}_{\mathtt{t}},\mathtt{a}) + \mathcal{C}_{\mathtt{t}}(f_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}},\mathtt{a})). \tag{2.1}$$

Recall that since we have a Deterministic Decision Process $f_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}},\mathtt{a}_{\mathtt{t}}) \in \mathcal{S}_{\mathtt{t}+1}$ is the next state if we take action $\mathtt{a}$ in $\mathtt{s}_{\mathtt{t}}$.

For the analysis, note that $\pi$ and $\pi'$ are identical until time $\mathtt{t}$, so they generate exactly the same distribution over paths. At time $\mathtt{t}$, $\pi'$ is define to minimize the cost to go from $\mathtt{s}_{\mathtt{t}}$, given that we follow $\pi$ from $\mathtt{t}+1$ to $\mathtt{T}$. Therefore the cost can only decrease. Formally, let $\mathbb{E}^{\pi}[\cdot]$ be the expectation with respect to policy $\pi$.

$$\begin{aligned}\mathbb{E}^{\pi}_{\mathtt{s}_{\mathtt{t}}}[\mathcal{C}_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}})] =&\mathbb{E}^{\pi}_{\mathtt{s}_{\mathtt{t}}}\mathbb{E}^{\pi}_{\mathtt{a}_{\mathtt{t}}}[\mathtt{c}(\mathtt{s}_{\mathtt{t}},\mathtt{a}_{\mathtt{t}}) + \mathcal{C}_{\mathtt{t}+1}(f_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}},\mathtt{a}_{\mathtt{t}}))]\\ \geq&\mathbb{E}^{\pi}_{\mathtt{s}_{\mathtt{t}}}\min_{\mathtt{a}_{\mathtt{t}}\in\mathcal{A}_{\mathtt{t}}}[\mathtt{c}(\mathtt{s}_{\mathtt{t}},\mathtt{a}_{\mathtt{t}}) + \mathcal{C}_{\mathtt{t}+1}(f_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}},\mathtt{a}_{\mathtt{t}}))]\\ =&\mathbb{E}^{\pi'}_{\mathtt{s}_{\mathtt{t}}}[\mathcal{C}_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}})]\end{aligned}$$

which completes the inductive proof. □

**Remark 2.5.** *The above proof extends very naturally for the case of a stochastic MDP, which implies that $f_{\mathtt{t}}$ is stochastic. The modification of the proof would simply take the expectation over $f_{\mathtt{t}}$ in Eq. 2.1.*

### 2.2.5 Optimal Control Policies

**Definition 2.2.** *A control policy $\pi \in \Pi_{MD}$ is called* ***optimal*** *if, for each initial state $\mathtt{s}_0$, it induces an optimal path $\mathtt{h}_{\mathtt{T}}$ from $\mathtt{s}_0$.*

An alternative definition can be given in terms of policies only. For that purpose, let $\mathtt{h}_{\mathtt{T}}(\pi;\mathtt{s}_0)$ denote the path induced by the policy $\pi$ from $\mathtt{s}_0$. For a given return functional $\mathcal{V}_{\mathtt{T}}(\mathtt{h}_{\mathtt{T}})$, denote $\mathcal{V}_{\mathtt{T}}(\pi;\mathtt{s}_0) = \mathcal{V}_{\mathtt{T}}(\mathtt{h}_{\mathtt{T}}(\pi;\mathtt{s}_0))$ That is, $\mathcal{V}_{\mathtt{T}}(\pi;\mathtt{s}_0)$ is the cumulative return for the path induced by $\pi$ from $\mathtt{s}_0$.

**Definition 2.3.** *A control policy $\pi \in \Pi_{MD}$ is called optimal if, for each initial state $\mathtt{s}_0$, it holds that $\mathcal{V}_{\mathtt{T}}(\pi;\mathtt{s}_0) \geq \mathcal{V}_{\mathtt{T}}(\tilde{\pi};\mathtt{s}_0)$ for any other policy $\tilde{\pi} \in \Pi_{MD}$.*

Equivalence of the two definitions can be easily established (exercise). An optimal policy is often denoted by $\pi^*$.

**The standard $\mathtt{T}$-stage finite-horizon planning problem:** Find a control policy $\pi$ for the $\mathtt{T}$-stage Finite Horizon problem that minimizes the cumulative cost (or maximizes the cumulative return) function.

**The naive approach to finding an optimal policy:** For finite models (i.e., finite state and action spaces), the number of feasible paths (or control policies) is finite. It is therefore possible, in principle, to enumerate all $\mathtt{T}$-stage paths, compute the cumulative return for each one, and choose the one which gives the largest return. Let us evaluate the number of different paths and control policies. Suppose for simplicity that number of states at each stage is the same: $|\mathcal{S}_{\mathtt{t}}| = n$, and similarly the number of actions at each state is the same: $|\mathcal{A}_{\mathtt{t}}(x)| = m$ (with $m \leq n$) . The number of feasible $\mathtt{T}$-stage paths for each initial state is seen to be $m^{\mathtt{T}}$. The number of different policies is $m^{n\mathtt{T}}$. For example, for a fairly small problem with $\mathtt{T} = n = m = 10$, we obtain $10^{10}$ paths for each initial state (and $10^{11}$ overall), and $10^{100}$ control policies. Clearly it is not computationally feasible to enumerate them all.

Fortunately, Dynamic Programming offers a drastic reduction of the computational complexity for this problem.

## 2.3 Finite Horizon Dynamic Programming

The Dynamic Programming (DP) algorithm breaks down the $\mathtt{T}$-stage finite-horizon problem into $\mathtt{T}$ sequential single-stage optimization problems. This results in dramatic improvement in computation efficiency.

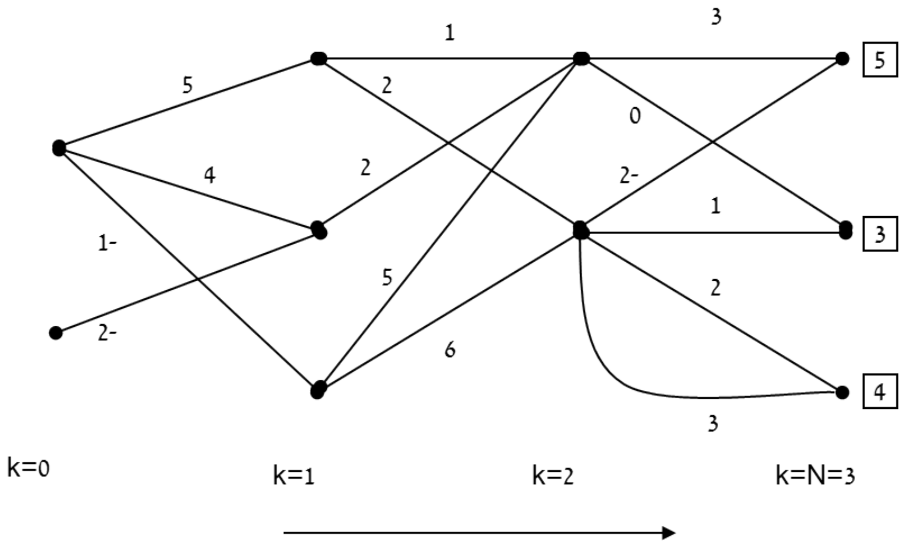
The DP technique for dynamic systems is based on a general observation called Bellman's Principle of Optimality. Essentially it states the following (for deterministic problems):

- **Any sub-path of an optimal path is itself an optimal path between its end point.**

To see why this should hold, consider a sub-path which is not optimal. We can replace it be an optimal sub-path, and improve the return.

Applying this principle recursively from the last stage backward, obtains the (backward) Dynamic Programming algorithm. Let us first illustrate the idea with following example.

**Example 2.3.** *Shortest path on a decision graph: Suppose we wish to find the shortest path (minimum cost path) from the initial node in* $\mathtt{T}$ *steps.*



*The boxed values are the terminal costs at stage* T*, the other number are the link costs. Using backward recursion, we may obtain that the minimal path costs from the two initial states are* 7 *and* 3*, as well as the optimal paths and an optimal policy.*

We can now describe the DP algorithm. Recall that we consider the dynamic system

$$\mathtt{s}_{\mathtt{t}+1} = f_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}), \quad \mathtt{t} = 0, 1, 2, \ldots, \mathtt{T} - 1$$

$$\mathtt{s}_{\mathtt{t}} \in \mathcal{S}_{\mathtt{t}}, \quad \mathtt{a}_{\mathtt{t}} \in \mathcal{A}_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}})$$

and we wish to maximize the cumulative return:

$$\mathcal{V}_{\mathtt{T}} = \sum_{\mathtt{t}=0}^{\mathtt{T}-1} \mathtt{r}_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) + \mathtt{r}_{\mathtt{T}}(\mathtt{s}_{\mathtt{T}})$$

The DP algorithm computes recursively a set of **value functions** $\mathcal{V}_{\mathtt{t}} : \mathcal{S}_{\mathtt{t}} \to \mathbb{R}$ , where $\mathcal{V}_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}})$ is the value of an optimal sub-path $\mathtt{h}_{\mathtt{t}:\mathtt{T}} = (\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}, \ldots, \mathtt{s}_{\mathtt{T}})$ that starts at $\mathtt{s}_{\mathtt{t}}$.

**Algorithm 2.1.** ***Finite-horizon Dynamic Programming***

1. *Initialize the value function:* $\mathcal{V}_{\mathtt{T}}(\mathtt{s}) = \mathtt{r}_{\mathtt{T}}(\mathtt{s})$, $\mathtt{s} \in \mathcal{S}_{\mathtt{T}}$.

2. *Backward recursion: For* $\mathtt{t} = \mathtt{T} - 1, \ldots, 0$, *compute*

$$\mathcal{V}_{\mathtt{t}}(\mathtt{s}) = \max_{\mathtt{a} \in \mathcal{A}_{\mathtt{t}}} \left\{ \mathtt{r}_{\mathtt{t}}(\mathtt{s}, \mathtt{a}) + \mathcal{V}_{\mathtt{t}+1}(f_{\mathtt{t}}(\mathtt{s}, \mathtt{a})) \right\}, \quad \mathtt{s} \in \mathcal{S}_{\mathtt{t}}.$$

3. *Optimal policy: Choose any control policy $\pi^* = (\pi^*_{\mathtt{t}})$ that satisfies:*

$$\pi^*_{\mathtt{t}}(\mathtt{s}) \in \arg\max_{\mathtt{a}\in\mathcal{A}_{\mathtt{t}}} \left\{ \mathtt{r}_{\mathtt{t}}(\mathtt{s},\mathtt{a}) + \mathcal{V}_{\mathtt{t}+1}(f_{\mathtt{t}}(\mathtt{s},\mathtt{a})) \right\}, \quad \mathtt{t} = 0,\ldots,\mathtt{T}-1.$$

**Proposition 2.3.** *The following holds for finite-horizon dynamic programming:*

1. *The control policy $\pi^*$ computed in Algorithm 2.1 is an optimal control policy for the* $\mathtt{T}$*-stage Finite Horizon problem.*

2. *$\mathcal{V}_0(\mathtt{s})$ is the optimal* $\mathtt{T}$*-stage return from initial state $\mathtt{s}_0 = \mathtt{s}$:*

$$\mathcal{V}_0(s) = \max_{\pi} V_0^{\pi}(\mathtt{s}), \quad \forall \mathtt{s} \in \mathcal{S}_0,$$

*where $V_0^{\pi}(\mathtt{s})$ is the expected return of policy $\pi$ when started at state $\mathtt{s}$.*

*Proof.* We show that the computed policy $\pi^*$ is optimal and its return from time $\mathtt{t}$ is $\mathcal{V}_{\mathtt{t}}$. We will establish the following inductive claim:

*For any time $\mathtt{t}$ and any state $\mathtt{s}$, the path from $\mathtt{s}$ define by $\pi^*$ is the maximum return path of length $\mathtt{T}-\mathtt{t}$. The value of $\mathcal{V}_{\mathtt{t}}(\mathtt{s})$ is the maximum return from $\mathtt{s}$.*

The proof is by a backward induction. For the basis of the induction we have: $\mathtt{t} = \mathtt{T}$, and the inductive claim follows from the initialization.

Assume the inductive claim holds for $\mathtt{t}$ prove for $\mathtt{t}+1$. For contradiction assume there is a higher return path from $\mathtt{s}$. Let the path generated $\pi^*$ be $P = (\mathtt{s}, \mathtt{s}^*_{\mathtt{T}-\mathtt{t}}, \ldots, \mathtt{s}^*_{\mathtt{T}})$. Let $P_1 = (\mathtt{s}, \mathtt{s}_{\mathtt{T}-\mathtt{t}}, \ldots, \mathtt{s}_{\mathtt{T}})$ be an alternative path. Let $P_2 = (\mathtt{s}, \mathtt{s}_{\mathtt{T}-\mathtt{t}}, \mathtt{s}'_{\mathtt{T}-\mathtt{t}+1}, \ldots, \mathtt{s}'_{\mathtt{T}})$ be the path generated by following $\pi^*$ from $\mathtt{s}_{\mathtt{T}-\mathtt{t}}$. Since $P_1$ and $P_2$ are identical except for the last $\mathtt{t}$ stages, we can use the inductive hypothesis, which implies that $\mathcal{V}(P_1) \leq \mathcal{V}(P_2)$. From the definition of $\pi^*$ we have that $\mathcal{V}(P_2) \leq \mathcal{V}(P)$. Hence, $\mathcal{V}(P_1) \leq \mathcal{V}(P_2) \leq \mathcal{V}(P)$, which completes the proof of the inductive hypothesis. □

Let us make the following observations:

1. The algorithm involves visiting each state exactly once, proceeding backward in time. For each time instant (or stage) $\mathtt{t}$, the value function $\mathcal{V}_{\mathtt{t}}(\mathtt{s})$ is computed for all states $\mathtt{s} \in \mathcal{S}_{\mathtt{t}}$ before proceeding to stage $\mathtt{t}-1$.

2. The *backward induction* step of Algorithm 2.1, along with similar equations in the theory of DP, is called **Bellman's equation**.

3. Computational complexity: There is a total of $n\mathtt{T}$ states (excluding the final one), and in each we need $m$ computations. Hence, the number of required calculations is $mn\mathtt{T}$. For the example above with $m = n = \mathtt{T} = 10$, we need $O(10^3)$ calculations.

4. A similar algorithm that proceeds forward in time (from $\mathtt{t} = 0$ to $\mathtt{t} = \mathtt{T}$) can be devised. We note that this will not be possible for stochastic systems (i.e., the stochastic MDP model).

5. The celebrated **Viterbi algorithm** is an important instance of finite-horizon DP. The algorithm essentially finds the most likely sequence of states in a Markov chain $(\mathtt{s_t})$ that is partially (or noisily) observed. The algorithm was introduced in 1967 for decoding convolution codes over noisy digital communication links. It has found extensive applications in communications, and is a basic computational tool in Hidden Markov Models (HMMs), a popular statistical model that is used extensively in speech recognition and bioinformatics, among other areas.

## 2.4 Shortest Paths

We can formulate a DDP problems similar to shortest path problems. Given a directed graph $G(V, E)$, there is a set of goal states $\mathcal{S}_G$, and the goal is to reach one of the goal states. Formally, when we reach a goal state we stay there and have a zero cost. For such a DDP the optimal policy would be to compute a shortest path to one of the goal states.

There is a variety of algorithms for computing shortest paths.

1. Bellman-Ford: handles also negative weights, but assumes no negative cycle. Runs in time $O(|V| \cdot |E|)$.

2. Dijkstra: Assumes non-negative weights. Runs in time $O(|V| \log |E| + |E|)$

## 2.5 Linear Programming for Finite Horizon

In this section we will use linear programming to derive the optimal policy. We will see that both the primal and dual program will play an important part in defining the optimal policy. We will fix an initial state $\mathtt{s}_0$ and compute the optimal policy for it.

We will start with the primal linear program, which will compute the optimal policy. For each time $\mathtt{t}$, state $\mathtt{s}$ and action $\mathtt{a}$ we will have a variable $x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) \in \{0,1\}$ that will indicate by 1 that at time $\mathtt{t}$ we are at state $\mathtt{s}$ and perform action $\mathtt{a}$. For the terminal states $\mathtt{s}$ we will have a variable $x_{\mathtt{T}}(\mathtt{s}) \in \{0,1\}$ that will indicate whether we terminate at state $\mathtt{s}$.

Our main constraint will be a flow constraint, stating that if we reach state $\mathtt{s}$ at time $\mathtt{t}-1$ then we exit it at time $\mathtt{t}$. Formally,

$$\sum_{\mathtt{a}} x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) = \sum_{\mathtt{s}',\mathtt{a}':f_{\mathtt{t}-1}(\mathtt{s}',\mathtt{a}')=\mathtt{s}} x_{\mathtt{t}-1}(\mathtt{s}',\mathtt{a}').$$

and for terminal states simply

$$x_{\mathtt{T}}(\mathtt{s}) = \sum_{\mathtt{s}',\mathtt{a}':f_{\mathtt{T}-1}(\mathtt{s}',\mathtt{a}')=\mathtt{s}} x_{\mathtt{T}-1}(\mathtt{s}',\mathtt{a}')$$

The return, which we would like to maximize, would be

$$\sum_{\mathtt{t},\mathtt{s},\mathtt{a}} \mathtt{r}_{\mathtt{t}}(\mathtt{s},\mathtt{a}) x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) + \sum_{\mathtt{s}} \mathtt{r}_{\mathtt{T}}(\mathtt{s}) x_{\mathtt{T}}(\mathtt{s})$$

To have a linear program we will also need to relax the constraints of $\{0,1\}$ to $[0,1]$

The resulting linear program is the following.

$$\max_{x_{\mathtt{t}}(\mathtt{s},\mathtt{a}),x_{\mathtt{T}}(\mathtt{s})} \quad \sum_{\mathtt{t},\mathtt{s},\mathtt{a}} \mathtt{r}_{\mathtt{t}}(\mathtt{s},\mathtt{a}) x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) + \sum_{\mathtt{s}} \mathtt{r}_{\mathtt{T}}(\mathtt{s}) x_{\mathtt{T}}(\mathtt{s})$$

such that

$$\sum_{\mathtt{a}} x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) = \sum_{\mathtt{s}',\mathtt{a}':f_{\mathtt{t}-1}(\mathtt{s}',\mathtt{a}')=\mathtt{s}} x_{\mathtt{t}-1}(\mathtt{s}',\mathtt{a}'). \qquad \forall \mathtt{s} \in \mathcal{S}_{\mathtt{t}}, \mathtt{t} \in \mathbb{T}$$

$$x_{\mathtt{T}}(\mathtt{s}) = \sum_{\mathtt{s}',\mathtt{a}':f_{\mathtt{T}-1}(\mathtt{s}',\mathtt{a}')=\mathtt{s}} x_{\mathtt{T}-1}(\mathtt{s}',\mathtt{a}') \qquad \forall \mathtt{s} \in \mathcal{S}_{\mathtt{T}}$$

$$x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) \geq 0 \qquad \forall \mathtt{s} \in \mathcal{S}_{\mathtt{t}}, \mathtt{a} \in \mathcal{A}, \mathtt{t} \in \{0,\ldots,\mathtt{T}-1\}$$

$$\sum_{\mathtt{a}} x_0(\mathtt{s}_0,\mathtt{a}) = 1$$

$$x_0(\mathtt{s},\mathtt{a}) = 0, \qquad \forall \mathtt{s} \in \mathcal{S}_0, \mathtt{s} \neq \mathtt{s}_0$$

Note that we (implicitly) have $\sum_{\mathtt{a}\in\mathcal{A},\mathtt{s}\in\mathcal{S}_\mathtt{t}} x_\mathtt{t}(\mathtt{s},\mathtt{a}) = 1$, and we can prove it by induction on $\mathtt{t}$.

Given the primal linear program we can derive the dual linear program.

$$\begin{aligned}
&\min_{z_\mathtt{t}(\mathtt{s})}\ z_0(\mathtt{s}_0)\\
&\text{such that}\\
&\qquad z_\mathtt{T}(\mathtt{s}) = \mathtt{r}_\mathtt{T}(\mathtt{s}) && \forall \mathtt{s}\in\mathcal{S}_\mathtt{t}\\
&\qquad z_\mathtt{t}(\mathtt{s}) \ge \mathtt{r}_\mathtt{t}(\mathtt{s},\mathtt{a}) + z_{\mathtt{t}+1}(f_\mathtt{t}(\mathtt{s},\mathtt{a})), && \forall \mathtt{s}\in\mathcal{S}_\mathtt{t}, \mathtt{a}\in\mathcal{A}, \mathtt{t}\in\mathbb{T},\\
&\qquad\qquad .
\end{aligned}$$

One can identify the dual random variables $z_\mathtt{t}(\mathtt{s})$ with the optimal vale function $\mathcal{V}_\mathtt{t}(\mathtt{s})$. At the optimal solution of the dual linear program one can show that we have

$$z_\mathtt{t}(\mathtt{s}) = \max_{\mathtt{a}}\left\{\mathtt{r}_\mathtt{t}(\mathtt{s},\mathtt{a}) + z_{\mathtt{t}+1}(f_\mathtt{t}(\mathtt{s},\mathtt{a}))\right\}, \qquad \forall \mathtt{s}\in\mathcal{S}_\mathtt{t}, \mathtt{t}\in\mathbb{T},$$

which is the familiar Bellman optimality equations.

## 2.6 Average cost criteria

The average cost criteria considers the limit of the average costs. Formally:

$$\mathcal{C}^{\pi}_{avg} = \lim_{\mathtt{T}\to\infty}\frac{1}{\mathtt{T}}\sum_{\mathtt{t}=0}^{\mathtt{T}-1}\mathtt{c}_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})$$

where the trajectory is generated using $\pi$. The aim is to minimize $\mathbb{E}[\mathcal{C}^{\pi}_{avg}]$. This implies that any finite prefix has no influence of the final average cost, since its influence vanishes as $\mathtt{T}$ goes to infinity.

For a deterministic stationary policy, the policy converges to a simple cycle, and the average cost is the average cost of the edges on the cycle. (Recall, we are considering only DDP.)

Given a directed graph $G(V,E)$, let $\Omega$ be the collection of all cycles in $G(V,E)$. For each cycle $\omega = (v_1,\ldots,v_k)$, we define $c(\omega) = \sum_{i=1}^{k} c(v_i, v_{i+1})$, where $(v_i, v_{i+1})$ is the $i$-th edge in the cycle $\omega$. Let $\mu(\omega) = \frac{c(\omega}{k}$. The *min average cost cycle* is

$$\mu^* = \min_{\omega\in\Omega}\mu(\omega)$$

We show that the min average cost cycle is the optimal policy.

**Theorem 2.4.** *For any Deterministic Decision Process (DPP) the optimal average costs is $\mu^*$, and an optimal policy is $\pi_\omega$ that cycles around a simple cycle of average cost $\mu^*$.*

*Proof.* Let $\omega$ be a cycle of average cost $\mu^*$. Let $\pi_\omega$ be a deterministic stationary policy that first reaches $\omega$ and then cycles in $\omega$. Clearly, $\mathcal{C}_{avg}^{\pi_\omega} = \mu^*$.

We show that for any policy $\pi$ (possibly in $\Pi_{HS}$) we have that $\mathbb{E}[\mathcal{C}_{avg}^{\pi}] \geq \mu^*$. For contradiction assume that there is a policy $\pi'$ that has average cost $\mu^* - \varepsilon$. Consider a sufficiently long run of length $\mathtt{T}$ of $\pi'$, and fix any realization $\theta$ of it. We will show that the cumulative cost $\mathcal{C}(\theta) \geq (\mathtt{T} - n)\mu^*$, which implies that $\mathbb{E}[\mathcal{C}_{avg}^{\pi}] \geq \mu^* - n\mu^*/\mathtt{T}$.

Given $\theta$, consider the first simple cycle $\omega$ in $\theta$. The average cost of $\omega$ is $\mu(\omega) \geq \mu^*$. Delete $\omega$ from $\theta$, reducing the number of edges by $|\omega|$ and the cumulative cost by $\mu(\omega)|\omega|$. We continue the process until there is no remaining cycles, which implies that we have at most $|V| = n$ nodes remaining. Therefore, the costs of $\omega$ was at least $(\mathtt{T} - n)\mu^*$. This implies that the average cost of $\theta$ is at least $\mathcal{C}(\theta) \geq (1 - \frac{n}{\mathtt{T}})\mu^*$. For $\epsilon > \mu^* n/\mathtt{T}$ we have a contradiction. $\square$

Next we develop an algorithm for computing the minimum average cost cycle, which implies an optimal policy for DDP for average costs. The input is a directed graph $G(V, E)$ with cost $\mathtt{c} : E \to \mathbb{R}$.

We first give a characterization of $\mu^*$. Set a root $r \in V$. Let $F_k(v)$ be paths of length $k$ from $r$ to $v$. Let $d_k(v) = \min_{p \in F_k(v)} \mathtt{c}(p)$, where if $F_k(v) = \emptyset$ then $d_k(v) = \infty$. The following theorem (due to [R. Karp, A characterization of the minimum cycle mean in digraph, Discrete mathematics, 1978]) gives a characterization of $\mu^*$.

**Theorem 2.5.**

$$\mu^* = \min_{v \in S} \max_{0 \leq k \leq n-1} \frac{d_n(v) - d_k(v)}{n - k} ,$$

*where we define $\infty - \infty$ as $\infty$.*

*Proof.* We have two cases, $\mu^* = 0$ and $\mu^* > 0$. We assume that the graph has no negative cycle (we can guarantee this by adding a large number $M$ to all the weights).

We start with $\mu^* = 0$. This implies that we have in $G(V, E)$ a cycle of weight zero, but no negative cycle. For the theorem it is sufficient to show that

$$\min_{v \in S} \max_{0 \leq k \leq n-1} \{d_n(v) - d_k(v)\} = 0.$$

For every node $v \in V$ there is a path of length $k \in [0, n-1]$ of cost $d(v)$, the cost of the shortest path from $r$ to $v$. This implies that

$$\max_{0 \leq k \leq n-1} \{d_n(v) - d_k(v)\} = d_n(v) - d(v) \geq 0$$

We need to show that for some $v \in V$ we have $d_n(v) = d(v)$, which implies that $\min_{v\in S}\{d_n(v) - d(v)\} = 0$.

Consider a cycle $\omega$ of cost $\mathcal{C}(\omega) = 0$ (there is one, since $\mu^* = 0$). Let $v$ be a node on the cycle $\omega$. Consider a path $P$ from $r$ to $v$ which then cycles around $\omega$ and has length at least $n$. The path $P$ is a shortest path to $v$ (although not necessarily simple). This implies that any sub-path of $P$ is also a shortest path. Let $P'$ be a sub-path of $P$ of length $n$ and let it end in $u \in V$. Path $P'$ is a shortest path to $u$, since it is a prefix of a shortest path $P$. This implies that the cost of $P'$ is $d(u)$. Since $P'$ is of length $n$, by construction, we have that $d_n(u) = d(u)$. Therefore, $\min_{v\in S}\{d_n(v) - d(v)\} = 0$, which completes the case that $\mu^* = 0$.

For $\mu^* > 0$ we subtract a constant $\Delta = \mu^*$ from al the costs in the graph. This implies that for the new costs we have a zero cycle and no negative cycle. We can now apply the previous case. It only remains to show that the formula changes by exactly $\Delta = \mu^*$.

Formally, for every edge $e \in E$ let $\mathsf{c}'(e) = \mathsf{c}(e) - \Delta$. For any path $p$ we have $\mathcal{C}'(p) = \mathcal{C}(p) - |p|\Delta$, and for any cycle $\omega$ we have $\mu'(\omega) = \mu(\omega) - \Delta$. This implies that for $\Delta = \mu^*$ we have a cycle of cost zero and no negative cycles. We now consider the formula,

$$\begin{aligned}
0 = (\mu')^* &= \min_{v\in V} \max_{0\le k\le n-1} \{\frac{d'_n(v) - d'_k(v)}{n-k}\} \\
&= \min_{v\in V} \max_{0\le k\le n-1} \{\frac{d_n(v) - n\Delta - d_k(v) + k\Delta}{n-k}\} \\
&= \min_{v\in V} \max_{0\le k\le n-1} \{\frac{d_n(v) - d_k(v)}{n-k} - \Delta\} \\
&= \min_{v\in V} \max_{0\le k\le n-1} \{\frac{d_n(v) - d_k(v)}{n-k}\} - \Delta
\end{aligned}$$

Therefore we have

$$\mu^* = \Delta = \min_{v\in V} \max_{0\le k\le n-1} \{\frac{d_n(v) - d_{\mathtt{t}}(v)}{n-k}\}$$

□

We would like now to recover the minimum average cost cycle. The basic idea is to recover the cycle from the minimizing vertices in the formula, but some care need to be taken. It is true that some minimizing pair $(v, k)$ the path of length $n$ from $r$ to $v$ has a cycle of length $n-k$ which is the suffix of the path. The solution is that for the path $p$, from $r$ to $v$ of length $n$, any simple cycle is a minimum average cost cycle. (See, [“A note of finding minimum mean cycle”, Mmamu Chaturvedi and Ross M. McConnell, IPL 2017]).

The running time of computing the minimum average cost cycle is $O(|V| \cdot |E|)$.

## 2.7 Historical Notes:

- Dynamic Programming was popularized in the 1950’s and 1960’s by **Richard Bellman** (1920-1984), an American mathematician. Bellman, who coined the term Dynamic Programming, formulated the general theory and proposed numerous applications in control and operations research.

- **Andrew Viterbi** (born 1935) is an American professor of electric engineer, a pioneer in the field of digital communications, and co-founder of Qualcomm Inc. (together with Irwin Jacobs). He has had close ties with the Technion, and has made many contributions to our department.

- **Richard Karp** (born 1935) is an American professor of computer science, known for his contributions to the theory of computing.

# Chapter 3

# Markov Chains

We start by considering a stochastic dynamics model which does not have any actions, so that we can concentrate on the dynamics.

A Markov chain $\{X_{\mathtt{t}},\ \mathtt{t}=0,1,2,\ldots\}$, with $X_{\mathtt{t}} \in X$, is a discrete-time stochastic process, over a finite or countable state-space $X$, that satisfies the following Markov property:

$$\mathcal{P}(X_{\mathtt{t}+1}=j|X_{\mathtt{t}}=i, X_{\mathtt{t}-1}, \ldots X_0) = \mathcal{P}(X_{\mathtt{t}+1}=j|X_{\mathtt{t}}=i).$$

We focus on time-homogeneous Markov chains, where

$$\mathcal{P}(X_{\mathtt{t}+1}=j|X_{\mathtt{t}}=i) = \mathcal{P}(X_1=j|X_0=i) \overset{\Delta}{=} p_{i,j}.$$

The $p_{i,j}$'s are the transition probabilities, which satisfy $p_{i,j} \geq 0$, and for each $i \in X$ we have $\sum_{j\in X} p_{i,j} = 1$, namely, $\{p_{i,j} : j \in X\}$ is a distribution on the next state following state $i$. The matrix $P = (p_{i,j})$ is the transition matrix. The matrix is row-stochastic (each row sums to 1 and all entries non-negative).

Given the initial distribution $p_0$ of $X_0$, namely $\mathcal{P}(X_0=i) = p_0(i)$, we obtain the finite-dimensional distributions:

$$\mathcal{P}(X_0=i_0,\ldots,X_{\mathtt{t}}=i_{\mathtt{t}}) = p_0(i_0)p_{i_0,i_1}\cdot\ldots\cdot p_{i_{\mathtt{t}-1},i_{\mathtt{t}}}.$$

Define $p_{i,j}^{(m)} = \mathcal{P}(X_m=j|X_0=i)$, the $m$-step transition probabilities. It is easy to verify that $p_{i,j}^{(m)} = [P^m]_{ij}$ (here $P^m$ is the $m$-th power of the matrix $P$).

**Example 3.1.** [1] *Consider the following two state Markov chain, with transition probability $P$ and initial distribution $p_0$, as follows:*

$$P = \begin{pmatrix} 0.4 & 0.6 \\ 0.2 & 0.8 \end{pmatrix} \qquad p_0 = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}$$

[1] Simple MC, two states, running example.

*Initially, we have both states equally likely. After one step, the distribution of states is* $p_1 = p_0 P = (0.3\ ,\ 0.7)$. *After two steps we have* $p_2 = p_1 P = p_0 P^2 = (0.26\ ,\ 0.74)$. *The limit of this sequence would be* $p_\infty = (0.25\ ,\ 0.75)$, *which is called the* steady state distribution, *and would be discussed later.*

**State classification:**

- State $j$ is *accessible* from $i$ (denoted by $i \to j$) if $p^{(m)}_{i,j} > 0$ for some $m \geq 1$.
  For a finite $X$ we can compute the accessibility property as follows. Construct a directed graph $G(X, E)$ where the vertices are the states $X$ and there is a directed edge $(i, j)$ if $p_{i,j} > 0$. State $j$ is accessible from state $i$ iff there exists a directed path in $G(X, E)$ from $i$ to $j$.

- States $i$ and $j$ are *communicating* states (or communicate) if $i \to j$ and $j \to i$.
  For a finite $X$, this implies that in $G(X, E)$ there is both a directed path from $i$ to $j$ and from $j$ to $i$.

- A *communicating class* (or just class) is a maximal collection of states that communicate.
  For a finite $X$, this implies that in $G(X, E)$ we have $i$ and $j$ in the same strongly connected component of the graph. (A strongly connected component has a directed path between any pair of vertices.)

- The Markov chain is *irreducible* if all states belong to a single class (i.e., all states communicate with each other).
  For a finite $X$, this implies that in $G(X, E)$ is strongly connected.

- State $i$ has a *period* $d_i = GCD\{m \geq 1 : p^{(m)}_{i,i} > 0\}$, where $GCD$ is the greatest common divisor. A state is aperiodic if $d_i = 1$.
  State $i$ is periodic with period $d_i \geq 2$ if $p^{(m)}_{i,i} = 0$ for $m \pmod{d_i} \neq 0$ and for any $m$ such that $p^{(m)}_{i,i} > 0$ we have $m \pmod{d_i} = 0$.
  If a state $i$ is aperiodic, then there exists an integer $m_0$ such that for any $m \geq m_0$ we have $p^{(m)}_{i,i} > 0$.

- Periodicity is a class property: all states in the same class have the same period. Specifically, if some state is *a-periodic*, then all states in the class are a-periodic.

**Claim 3.1.** *For any two states* $i$ *and* $j$ *with periods* $d_i$ *and* $d_j$*, in the same communicating class, we have* $d_i = d_j$.

*Proof.* For contradiction, assume that $d_j \pmod{d_i} \neq 0$. Since they are in the same communicating class, we have a trajectory from $i$ to $j$ of length $m_{i,j}$ and from $j$ to $i$ of length $m_{j,i}$. This implies that $(m_{i,j} + m_{j,i}) \pmod{d_i} = 0$. Now, there is a trajectory (which is a cycle) of length $m_{j,j}$ from $j$ back to $j$ such that $m_{j,j} \pmod{d_i} \neq 0$ (otherwise $d_i$ divides the period of $j$). Consider the path from $i$ to itself of length $m_{i,j} + m_{j,j} + m_{j,i}$. We have that $(m_{ij} + m_{jj} + m_{ji}) \pmod{d_i} = m_{jj} \pmod{d_i} \neq 0$. This is a contradiction to the definition of $d_i$. Therefore, $d_j \pmod{d_i} = 0$ and similarly $d_i \pmod{d_j} = 0$, which implies that $d_i = d_j$. □

The claim shows that periodicity is a class property, and all the states in a class have the same period.

**Example 3.2.** *Add figures explaining the definitions.*

**Recurrence:**

- State $i$ is *recurrent* if $\mathcal{P}(X_{\mathtt{t}} = i \text{ for some } t \geq 1 | X_0 = i) = 1$. Otherwise, state $i$ is *transient*.
  We can relate the state property of recurrent and transient to the expected number of returns to a state.

  **Claim 3.2.** *State $i$ is transient iff $\sum_{m=1}^{\infty} p_{i,i}^{(m)} < \infty$.*

  *Proof.* Assume that state $i$ is transient. Let $q_i = \mathcal{P}(X_{\mathtt{t}} = i \text{ for some } t \geq 1 | X_0 = i)$. Since state $i$ is transient we have $q_i < 1$. Let $Z_i$ be the number of times the trajectory returns to state $i$. Note that $Z_i$ is geometrically distributed with parameter $q_i$, namely $\Pr[Z_i = k] = q_i^k(1 - q_i)$. Therefore the expected number of returns to state $i$ is $1/(1 - q_i)$ and is finite. The expected number of returns to state $i$ is equivalently $\sum_{m=1}^{\infty} p_{i,i}^{(m)}$, and hence if a state is transient we have $\sum_{m=1}^{\infty} p_{i,i}^{(m)} < \infty$.

  For the other direction, assume that $\sum_{m=1}^{\infty} p_{i,i}^{(m)} < \infty$. This implies that there is an $m_0$ such that $\sum_{m=m_0}^{\infty} p_{i,i}^{(m)} < 1/2$. Consider the probability of returning to $i$ within $m_0$ stages. This implies that $\mathcal{P}(X_{\mathtt{t}} = i \text{ for some } t \geq m_0 | X_0 = i) < 1/2$. Now consider the probability $q_i' = \mathcal{P}(X_{\mathtt{t}} = i \text{ for some } m_0 \geq t \geq 1 | X_0 = i)$. If $q_i' < 1$, this implies that $\mathcal{P}(X_{\mathtt{t}} = i \text{ for some } t \geq 1 | X_0 = i) < q_i' + (1 - q_i')/2 = (1 + q_i')/2 < 1$, which implies that state $i$ is transient. If $q_i' = 1$, this implies that after at most $m_0$ stages we are guarantee to return to $i$, hence the expected number of return to state $i$ is infinite, i.e., $\sum_{m=1}^{\infty} p_{i,i}^{(m)} = \infty$. This is in contradiction to the assumption that $\sum_{m=1}^{\infty} p_{i,i}^{(m)} < \infty$. $\square$

- Recurrence is a class property.
  To see this consider two states $i$ and $j$ such that $j$ is accessible from $i$ and $i$ is recurrent. Since $i$ is recurrant $\sum_{m=1}^{\infty} p_{i,i}^{(m)} = \infty$. Since $j$ is accessible from $i$ there is a $k$ such that $p_{i,j}^{(k)} > 0$. Since $i$ is recurrent, there exists a $k'$ such that $p_{j,i}^{(k')} > 0$. We can lower bound $\sum_{m=1}^{\infty} p_{j,j}^{(m)}$ by $\sum_{m=1}^{\infty} p_{j,i}^{(k')} p_{i,i}^{(m)} p_{i,j}^{(k)} = \infty$. Therefore we showed that state $j$ is recurrent.

- If states $i$ and $j$ are in the same recurrent (communicating) class, then state $j$ is (eventually) reached from state $i$ with probability 1: $\mathcal{P}(X_{\mathtt{t}} = j \text{ for some t} \geq 1 | X_0 = i) = 1$.
  This follows from the fact that both states occur infinitely often, with probability 1.

- Let $T_i$ be the *return time* to state $i$ (number of stages required for $(X_{\mathtt{t}})$ starting from state $i$ to the first return to $i$). If $i$ is a recurrent state, then $T_i < \infty$ w.p. 1.
  Since otherwise, there is a positive probability that we never return to state $i$, and hence state $i$ is not recurrent.

- State $i$ is *positive recurrent* if $\mathbb{E}(T_i) < \infty$, and null recurrent if $\mathbb{E}(T_i) = \infty$. If the state space $X$ is finite, all recurrent states are positive recurrent.
  This follows since the set of states that are null recurrent cannot have transitions from positive recurrent states and cannot have transition to transient states. If the chain never leaves the set of null recurrent states, then some state would have a return time which is at most the size of the set. If there is a positive probability of leaving the set (and never returning) then the states are transient.

**Example 3.3. Random walk** *Consider the following Markov chain over the integers. The states are the integers. The initial states is* $0$. *At each state* $i$, *with probability* $1/2$ *we move to* $i+1$ *and with probability* $1/2$ *to* $i-1$. *Namely,* $p_{i,i+1} = 1/2$, $p_{i,i-1} = 1/2$, *and* $p_{i,j} = 0$ *for* $j \notin \{i-1, i+1\}$. *We will show that* $T_i$ *is finite with probability* $1$ *and* $\mathbb{E}[T_i] = \infty$. *This implies that all the states are null recurrent.*

*To compute* $\mathbb{E}[T_i]$ *consider what happens in one and two steps. Let* $Z_{i,i-1}$ *be the time to move from* $i$ *to* $i-1$. *Note that we have*

$$\mathbb{E}[T_i] = 1 + 0.5\mathbb{E}[Z_{i+1,i}] + 0.5\mathbb{E}[Z_{i-1,i}] = 1 + \mathbb{E}[Z_{1,0}],$$

*since, due to symmetry,* $\mathbb{E}[Z_{i,i+1}] = \mathbb{E}[Z_{i+1,i}] = \mathbb{E}[Z_{1,0}]$.

*In two steps we are either back to* $i$*, or at state* $i+2$ *or* $i-2$*. For* $\mathbb{E}[T_i]$ *we have that*

$$\begin{aligned}
\mathbb{E}[T_i] =& 2 + \frac{1}{4}\mathbb{E}[Z_{i+2,i}] + \frac{1}{4}\mathbb{E}[Z_{i-2,i}] \\
=& 2 + \frac{1}{4}(\mathbb{E}[Z_{i+2,i+1}] + \mathbb{E}[Z_{i+1,i}]) + \frac{1}{4}(\mathbb{E}[Z_{i-2,i-1}] + \mathbb{E}[Z_{i-1,i}]) \\
=& 2 + E[Z_{1,0}]
\end{aligned}$$

*This implies that we have*

$$\mathbb{E}[T_i] = 1 + \mathbb{E}[Z_{1,0}] = 2 + E[Z_{1,0}]$$

*Clearly, there is no finite value for* $\mathbb{E}[Z_{1,0}]$ *which will satisfy the equation, which implies* $\mathbb{E}[Z_{1,0}] = \infty$*, and hence* $\mathbb{E}[T_i] = \infty$.

*To show that* $0$ *is a recurrent state, note that the probability that at time* $2k$ *we are at state* $0$ *is exactly* $p_{0,0}^{(2k)} = \binom{2k}{k}2^{-2k} \approx \frac{c}{\sqrt{k}}$*, for some* $c > 0$*. This implies that*

$$\sum\nolimits_{m=1}^{\infty} p_{0,0}^{(m)} \approx \sum\nolimits_{m=1}^{\infty} \frac{c}{\sqrt{m}} = \infty$$

*and therefore state* $0$ *is recurrent. (By symmetry, this shows that all the states are recurrent).*

*Note that this Markov chain has a period of* $2$.

**Example 3.4. Random walk with jumps.** *Consider the following Markov chain over the integers. The states are the integers. The initial states is* $0$*. At each state* $i$*, with probability* $1/2$ *we move to* $i+1$ *and with probability* $1/2$ *we return to* $0$*. Namely,* $p_{i,i+1} = 1/2$*,* $p_{i,0} = 1/2$*, and* $p_{i,j} = 0$ *for* $j \notin \{0, i+1\}$*. We will show that* $\mathbb{E}[T_i] < \infty$ *(which implies that* $T_i$ *is finite with probability* $1$*).*

*From any state we return to* $0$ *with probability* $1/2$*, therefore* $\mathbb{E}[T_0] = 2$*. We will show that for state* $i$ *we have* $\mathbb{E}[T_i] \leq 2 + 2 \cdot 2^i$*. We will decompose* $T_i$ *to two parts. The first is the return to* $0$*, this part has expectation* $2$*. The second is to reach state* $i$ *from state* $0$*. Consider an epoch as the time between two visits to* $0$*. The probability that an epoch would reach* $i$ *is exactly* $2^{-i}$*. The expected time of an epoch is* $2$ *(the expected time to return to state* $0$*). The expected time to return to state* $0$*, given that we did not reach state* $i$ *is less than* $2$*. Therefore,* $\mathbb{E}[T_i] \leq 2 + 2 \cdot 2^i$.

*Note that this Markov chain is aperiodic.*

**Invariant Distribution:** The probability vector $\mu = (\mu_i)$ is an invariant distribution (or stationary distribution or steady state distribution) for the Markov chain if $\mu^\top P = \mu^\top$, namely

$$\mu_j = \sum_i \mu_i p_{i,j} \quad \forall j.$$

Clearly, if $X_{\mathtt{t}} \sim \mu$ then $X_{\mathtt{t}+1} \sim \mu$. If $X_0 \sim \mu$, then the Markov chain $(X_{\mathtt{t}})$ is a stationary stochastic process.

**Theorem 3.3.** *Let $(X_{\mathtt{t}})$ be an irreducible and a-periodic Markov chain over a finite state space $X$ with transition matrix $P$. Then there is a unique distribution $\mu$ such that $\mu^\top P = \mu^\top > 0$.*

*Proof.* Assume that $x$ is an eigenvector of $P$ with eigenvalue $\lambda$, i.e., we have $Px = \lambda x$. Since $P$ is a stochastic matrix, we have $\|Px\|_\infty \leq \|x\|_\infty$, which implies that $\lambda \leq 1$. Since the matrix $P$ is stochastic, $P\vec{1} = \vec{1}$, which implies that $P$ has a right eigenvalue of 1 and this is the maximal eigenvalue. Since the set of right and left eignevalues is identical, we conclude that there is $x$ such that $x^\top P x^\top$. Our first task is to show that there is such an $x$ with $x \geq 0$.

Since the Markov chain is irreducible and a-periodic, there is an integer $m$, such that $P^m$ has all the entries strictly positive. Namely, for any $i, j \in X$ we have $p_{i,j}^{(m)} > 0$.

We now show a general property of positive matrices (matrices where are the entries are strictly positive). Let $A = P^m$ be a positive matrix and $x$ an eigenvector of $A$ with eigenvalue 1. First, if $x$ has complex number then $Re(x)$ and $Im(x)$ are an eigenvectors of $A$ of eigenvalue 1 and one of them is non-zero. Therefore we can assume that $x \in \mathbb{R}^d$. We would like to show that there is an $x \geq 0$ such that $x^\top A = x^\top$. If $x \geq 0$ we are done. If $x \leq 0$ we can take $x' = -x$ and we are done. We need to show that $x$ cannot have both positive and negative entries.

For contradiction, assume that we have $x_k > 0$ and $x_{k'} < 0$. This implies that for any weight vector $w > 0$ we have $|x^\top w| < |x|^\top w$, where $|x|$ is point-wise absolute value. Therefore,

$$\sum_j |x_j| = \sum_j |\sum_i x_i P_{i,j}| < \sum_j \sum_i |x_i| P_{i,j} = \sum_i |x_i| \sum_j P_{i,j} = \sum_j |x_j|$$

where the first identity follows since $x$ is an eigenvector. The second since $P$ is strictly positive.The third is a change of order of summation. The last follows since $P$ is a stochastic matrix, so each row sums to 1. Clearly, we reached an contradiction, and therefore $x$ cannot have both positive and negative entries.

We have shown so far that there exists a $\mu$ such that $\mu^\top P = \mu^\top$ and $\mu \geq 0$. Since $A = P^m$ is strictly positive, then $\mu^\top = \mu^\top A > 0$.

To show the uniqueness of $\mu$, assume we have $x$ and $y$ such that $x^\top P = x^\top$ and $y^\top P = y^\top$ and $x \neq y$. Recall that we showed that in such a case both $x > 0$ and $y > 0$. Then there is a linear combination $z = ax + by$ such that for some $i$ we have $z_i = 0$. Since $z^\top P = z^\top$, we have showed that $z$ is strictly positive, i.e., $z > 0$, which is a contradiction. Therefore, $x = y$, and hence $\mu$ is unique. □

We define the average fraction that a state $j \in X$ occurs, given that we start with an initial state distribution $x_0$, as follows:

$$\pi_j^{(m)} = \frac{1}{m}\sum_{t=1}^{m} \mathbb{I}(X_t = j)$$

**Theorem 3.4.** *Let $(X_t)$ be an irreducible and a-periodic Markov chain over a finite state space $X$ with transition matrix $P$. Let $\mu$ be the stationary distribution of $P$. Then, for any $j \in X$ we have*

$$\mu_j = \lim_{m\to\infty} \mathbb{E}[\pi_j^{(m)}] = \frac{1}{\mathbb{E}[T_j]}$$

*Proof.* We have that

$$\mathbb{E}[\pi_j^{(m)}] = \mathbb{E}[\frac{1}{m}\sum_{t=1}^{m} \mathbb{I}(X_t = j)] = \frac{1}{m}\sum_{t=1}^{m} \Pr[X_t = j | X_0 = x_0] = \frac{1}{m}\sum_{t=1}^{m} x_0^\top P^t$$

Let $v_1, \ldots, v_n$ be the eigenvectors of $P$ with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$. By Theorem 3.3 that $v_1 = \mu$, the stationary distribution and $\lambda_1 = 1 > \lambda_i$ for $i \geq 2$. Rewrite $x_0 = \sum_i \alpha_i v_i$. Since $P^m$ is a stochastic matrix, $x_0^\top P^m$ is a distribution, and therefore $\lim_{m\to\infty} x_0^\top P^m = \mu$.

We will be interested in the limit $\pi_j = \lim_{m\to\infty} \pi_j^m$, and mainly in the expected value $\mathbb{E}[\pi_j]$. From the above we have that $\mathbb{E}[\pi_j] = \mu_j$.

A different way to express $\mathbb{E}[\pi_j]$ is using a variable time horizon, with a fixed number of occurrences of $j$. Let $T_{k,j}$ be the time between the $k$ and $k+1$ occurrence of state $j$. This implies that

$$\lim_{m\to\infty} \frac{1}{m}\sum_{t=1}^{m} \mathbb{I}(X_t = j) = \lim_{n\to\infty} \frac{n}{\sum_{k=1}^{n} T_{k,j}}$$

Note that the $T_{k,j}$ are i.i.d. and equivalent to $T_j$. By the law of large numbers we have that $\frac{1}{n}\sum_{k=1}^{n} T_{k,j}$ converges to $\mathbb{E}[T_i]$. Therefore,

$$\mathbb{E}[\pi_j] = \frac{1}{\mathbb{E}[T_j]}$$

□

We have established the following general theorem.

**Theorem 3.5 (Recurrence of finite Markov chains).** *Let $(X_{\mathtt{t}})$ be an irreducible, a-periodic Markov chain over a finite state space $X$. Then the following properties hold:*

1. *All states are* ***positive recurrent***
2. *There exists a* ***unique stationary distribution*** *$\mu^*$, where $\mu^*(i) = 1/\mathbb{E}[T_i]$.*
3. ***Convergence*** *to the stationary distribution:* $\lim_{t\to\infty} p_{i,j}^{(t)} = \mu_j \quad (\forall j)$
4. ***Ergodicity****: For any finite $f$:* $\lim_{t\to\infty} \frac{1}{t}\sum_{s=0}^{\mathtt{t}-1} f(X_s) = \sum_i \mu_i f(i) \overset{\Delta}{=} \pi \cdot f$.

For countable Markov chains, there are other possibilities.

**Theorem 3.6 (Countable Markov chains).** *Let $(X_{\mathtt{t}})$ be an irreducible and a-periodic Markov chain over a countable state space $X$. Then: Either (i) all states are positive recurrent, or (ii) all states are null recurrent, or (iii) all states are transient.*

*Proof.* Let $i$ is a positive recurrent state, then we will show that all states are positive recurrent. For any state $j$, since the Markov chain is irreducible, we have for some $m_1, m_2 \geq 0$ that $p_{j,i}^{(m_1)}, p_{i,j}^{(m_2)} > 0$. This implies that the return time to state $j$ is at most $\mathbb{E}[T_j] \leq 1/p_{j,i}^{(m_1)} + \mathbb{E}[T_i] + 1/p_{i,j}^{(m_2)}$, and hence $j$ is positive recurrent.

If there is no positive recurrent state, let $i$ be a null recurrent state, then we will show that all states are null recurrent. For any state $j$, since the Markov chain is irreducible, we have for some $m_1, m_2 \geq 0$ that $p_{j,i}^{(m_1)}, p_{i,j}^{(m_2)} > 0$. This implies that the probability to return time to state $j$ is at least $\sum_{m=0}^{\infty} p_{j,i}^{(m_1)}\ p_{i,i}^{(m)}\ p_{i,j}^{(m_2)} = \infty$, since we have $\sum_{m=0}^{\infty} p_{i,i}^{(m)} = \infty$, since $i$ is a recurrent state. This implies that $j$ is a recurrent state. Since there are no positive recurrent states, it has to be that $j$ is a null recurrent state.

If there are no positive or null recurrent states, then all states are transient. □

**Reversible Markov chains:** Suppose there exists a probability vector $\mu = (\mu_i)$ so that

$$\mu_i p_{i,j} = \mu_j p_{j,i}, \qquad i, j \in X. \tag{3.1}$$

It is then easy to verify by direct summation that $\mu$ is an invariant distribution for the Markov chain defined by $(p_{i,j})$. This follows since $\sum_i \mu_i p_{i,j} = \sum_i p_{i,j} \mu_j = \mu_j$. The equations (3.1) are called the *detailed balance equations.* A Markov chain that satisfies these equations is called reversible.

**Example 3.5** (**Discrete-time queue**). *Consider a discrete-time queue, with queue length $X_{\mathtt{t}} \in \mathbb{N}_0 = \{0, 1, 2, \ldots\}$. At time instant $t$, $A_{\mathtt{t}}$ new jobs arrive, and then up to $S_{\mathtt{t}}$ jobs can be served, so that*

$$X_{\mathtt{t+1}} = (X_{\mathtt{t}} + A_{\mathtt{t}} - S_{\mathtt{t}})^+.$$

*Suppose that $(S_{\mathtt{t}})$ is a sequence of i.i.d. RVs, and similarly $(A_{\mathtt{t}})$ is a sequence of i.i.d. RVs, with $(S_{\mathtt{t}})$, $(A_{\mathtt{t}})$ and $X_0$ mutually independent. It may then be seen that $(X_{\mathtt{t}},\ t \geq 0)$ is a Markov chain. Suppose further that each $S_{\mathtt{t}}$ is a Bernoulli RV with parameter $q$, namely $P(S_{\mathtt{t}} = 1) = q$, $P(S_{\mathtt{t}} = 0) = 1 - q$. Similarly, let $A_{\mathtt{t}}$ be a Bernoulli RV with parameter $p$. Then*

$$p_{i,j} = \begin{cases} p(1-q) & : \ j = i+1 \\ (1-p)(1-q) + pq & : \ j = i, \ \ i > 0 \\ (1-p)q & : \ j = i-1, \ \ i > 0 \\ (1-p) + pq & : \ j = i = 0 \\ 0 & : \ \text{otherwise} \end{cases}$$

*Denote $\lambda = p(1-q)$, $\eta = (1-p)q$, and $\rho = \lambda/\eta$. The detailed balance equations for this case are:*

$$\mu_i \lambda = \mu_{i+1} \eta, \qquad i \geq 0$$

*These equations have a solution with $\sum_i \mu_i = 1$ if and only if $\rho < 1$. The solution is $\mu_i = \mu_0 \rho^i$, with $\pi_0 = 1 - \rho$. This is therefore the stationary distribution of this queue.*

**Mixing time:** Add definition and motivation

# Chapter 4

# Markov Decision Processes and Finite Horizon

In Chapter 2 we considered multi-stage decision problems for *deterministic* systems. In many problems of interest, the system dynamics also involves *randomness*, which leads us to stochastic decision problems. In this chapter we introduce the basic model of Markov Decision Processes (MDP), which will be considered in the rest of the book, and discuss the finite horizon return.

## 4.1 Controlled Markov Chains

A Markov Decision Process consists of two main parts:

1. A controlled dynamic system, with stochastic evolution.
2. A performance objective to be optimized.

In this section we describe the first part, which is modeled as a controlled Markov chain.

Consider a controlled dynamic system, defined over:

- A discrete time axis $\mathbb{T} = \{0, 1, \ldots, \mathtt{T}-1\}$ (finite horizon), or $\mathbb{T} = \{0, 1, 2, \ldots\}$ (infinite horizon). To simplify the discussion we refer below to the infinite horizon case, which can always be "truncated" at $\mathtt{T}$ if needed.
- A finite state space $\mathcal{S}$, where $\mathcal{S}_\mathtt{t} \subset \mathcal{S}$ is the set of possible states at time $\mathtt{t} \in \mathbb{T}$.
- A finite action set $\mathcal{A}$, where $\mathcal{A}_\mathtt{t}(\mathtt{s}) \subset \mathcal{A}$ is the set of possible actions at time $\mathtt{t} \in \mathbb{T}$ and state $\mathtt{s} \in \mathcal{S}_\mathtt{t}$.

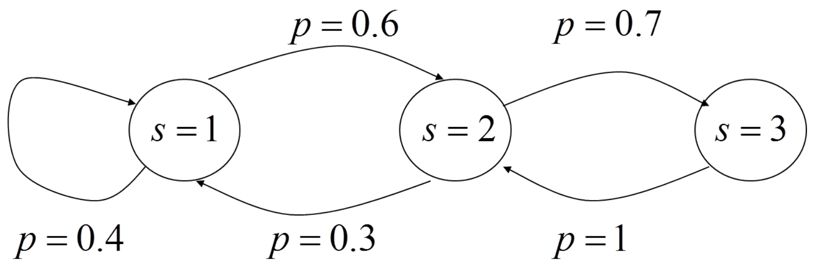


Figure 4.1: Markov chain

**State transition probabilities:**

- Suppose that at time $\mathtt{t}$ we are in state $\mathtt{s_t} = \mathtt{s}$, and choose an action $\mathtt{a_t} = \mathtt{a}$. The next state $\mathtt{s_{t+1}} = \mathtt{s}'$ is then determined randomly according to a probability distribution $p_{\mathtt{t}}(\cdot|\mathtt{s}, \mathtt{a})$ on $\mathcal{S}_{\mathtt{t+1}}$. That is,

$$\Pr(\mathtt{s_{t+1}} = \mathtt{s}'|\mathtt{s_t} = \mathtt{s}, \mathtt{a_t} = \mathtt{a}) = p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s}, \mathtt{a}), \qquad \mathtt{s}' \in \mathcal{S}_{\mathtt{t+1}}$$

- The probability $p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s}, \mathtt{a})$ is the *transition probability* from state $\mathtt{s}$ to state $\mathtt{s}'$ for a given action $\mathtt{a}$. We naturally require that $p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s}, \mathtt{a}) \geq 0$, and $\sum_{\mathtt{s}' \in \mathcal{S}_{\mathtt{t+1}}} p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s}, \mathtt{a}) = 1$ for all $\mathtt{s} \in \mathcal{S}_{\mathtt{t}}, \mathtt{a} \in \mathcal{A}_{\mathtt{t}}(\mathtt{s})$.

- Implicit in this definition is the controlled-Markov property:

$$\Pr(\mathtt{s_{t+1}} = \mathtt{s}'|\mathtt{s_t}, \mathtt{a_t}, \ldots, \mathtt{s}_0, \mathtt{a}_0) = \Pr(\mathtt{s_{t+1}} = \mathtt{s}'|\mathtt{s_t}, \mathtt{a_t})$$

- The set of probability distributions

$$P = \{p_{\mathtt{t}}(\cdot|\mathtt{s}, \mathtt{a}) \quad : \quad \mathtt{s} \in \mathcal{S}_{\mathtt{t}}, \mathtt{a} \in \mathcal{A}_{\mathtt{t}}(\mathtt{s}), \mathtt{t} \in \mathbb{T}\}$$

  is called the *transition law* or *transition kernel* of the controlled Markov process.

**Stationary Models:** The controlled Markov chain is called stationary or time-invariant if the transition probabilities do not depend on the time $\mathtt{t}$. That is:

$$\forall t, \quad p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s}, \mathtt{a}) \equiv p(\mathtt{s}'|\mathtt{s}, \mathtt{a}), \;\; \mathcal{S}_{\mathtt{t}} \equiv \mathcal{S}, \;\; \mathcal{A}_{\mathtt{t}}(\mathtt{s}) \equiv \mathcal{A}(\mathtt{s}).$$

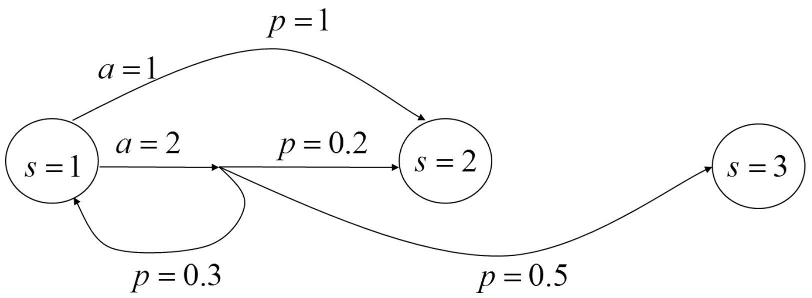


Figure 4.2: Controlled Markov chain

**Graphical Notation:** The state transition probabilities of a Markov chain are often illustrated via a state transition diagram, such as in Figure 4.1.

A graphical description of a controlled Markov chain is a bit more complicated because of the additional action variable. We obtain the diagram (drawn for state $\mathtt{s} = 1$ only, and for a given time $\mathtt{t}$) in Figure 4.2, reflecting the following transition probabilities:

$$p(\mathtt{s}' = 2|\mathtt{s} = 1, \mathtt{a} = 1) = 1$$

$$p(\mathtt{s}'|\mathtt{s} = 1, \mathtt{a} = 2) = \begin{cases} 0.3 & : & \mathtt{s}' = 1 \\ 0.2 & : & \mathtt{s}' = 2 \\ 0.5 & : & \mathtt{s}' = 3 \end{cases}$$

**State-equation notation:** The stochastic state dynamics can be equivalently defined in terms of a state equation of the form

$$\mathtt{s}_{\mathtt{t}+1} = f_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}, w_{\mathtt{t}}),$$

where $w_{\mathtt{t}}$ is a random variable (RV). If $(w_{\mathtt{t}})_{t\geq 0}$ is a sequence of independent RVs, and further each $w_{\mathtt{t}}$ is independent of the "past" $(\mathtt{s}_{\mathtt{t}-1}, \mathtt{a}_{\mathtt{t}-1}, \ldots \mathtt{s}_0)$, then $(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}})_{t\geq 0}$ is a controlled Markov process. For example, the state transition law of the last example can be written in this way, using $w_{\mathtt{t}} \in \{4, 5, 6\}$, with $p_w(4) = 0.3$, $p_w(5) = 0.2$, $p_w(6) = 0.5$ and, for $\mathtt{s}_{\mathtt{t}} = 1$:

$$f_{\mathtt{t}}(1, 1, w_{\mathtt{t}}) = 2$$
$$f_{\mathtt{t}}(1, 2, w_{\mathtt{t}}) = w_{\mathtt{t}} - 3$$

This state algebraic equation notation is especially useful for problems with continuous state space, but also for some models with discrete states. Equivalently, we can write

$$f_{\mathtt{t}}(1, 2, w_{\mathtt{t}}) = 1 \cdot \mathbb{I}[w_{\mathtt{t}} = 4] + 2 \cdot \mathbb{I}[w_{\mathtt{t}} = 5] + 3 \cdot \mathbb{I}[w_{\mathtt{t}} = 6],$$

where $\mathbb{I}[\cdot]$ is the indicator function.

Next we recall the definitions of control policies from Chapter 2.

## Control Policies

- A general or **history-dependent deterministic** control policy $\pi = (\pi_{\mathtt{t}})_{\mathtt{t}\in\mathbb{T}}$ is a mapping from each possible history $\mathtt{h_t} = (\mathtt{s}_0, \mathtt{a}_0, \ldots, \mathtt{s}_{\mathtt{t}-1}, \mathtt{a}_{\mathtt{t}-1}, \mathtt{s_t})$, and time $\mathtt{t} \in \mathbb{T}$, to an action $\mathtt{a_t} = \pi_{\mathtt{t}}(\mathtt{h_t}) \in \mathcal{A}_{\mathtt{t}}$. We denote the set of general policies by $\Pi_{HD}$.

- A **Markov deterministic** control policy $\pi$ is allowed to depend on the current state and time only: $\mathtt{a_t} = \pi_{\mathtt{t}}(\mathtt{s_t})$. We denote the set of Markov deterministic policies by $\Pi_{MD}$.

- For stationary models, we may define **stationary deterministic** control policies that depend on the current state alone. A stationary policy is defined by a single mapping $\pi : \mathcal{S} \to \mathcal{A}$, so that $\mathtt{a_t} = \pi(\mathtt{s_t})$ for all $\mathtt{t} \in \mathbb{T}$. We denote the set of stationary policies by $\Pi_{SD}$.

- Evidently, $\Pi_{HD} \supset \Pi_{MD} \supset \Pi_{SD}$.

## Randomized Control policies

- The control policies defined above specify deterministically the action to be taken at each stage. In some cases we want to allow for a random choice of action.

- A general randomized control policy assigns to each possible history $\mathtt{h_t}$ a probability distribution $\pi_{\mathtt{t}}(\cdot|\mathtt{h_t})$ over the action set $\mathcal{A}_{\mathtt{t}}$. That is, $\Pr(\mathtt{a_t} = \mathtt{a}|\mathtt{h_t}) = \pi_{\mathtt{t}}(\mathtt{a}|\mathtt{h_t})$. We denote the set of history-dependent stochastic policies by $\Pi_{HS}$.

- Similarly, we can define the set $\Pi_{MS}$ of Markov stochastic control policies, where $\pi_{\mathtt{t}}(\cdot|\mathtt{h_t})$ is replaced by $\pi_{\mathtt{t}}(\cdot|\mathtt{s_t})$, and the set $\Pi_{SS}$ of stationary stochastic control policies, where $\pi_{\mathtt{t}}(\cdot|\mathtt{s_t})$ is replaced by $\pi(\cdot|\mathtt{s_t})$, namely the policy is independent of the time.

- Note that the set $\Pi_{HS}$ includes all other policy sets as special cases.

**The Induced Stochastic Process** Let $p_0 = \{p_0(\mathtt{s}), \mathtt{s} \in \mathcal{S}_0\}$ be a probability distribution for the initial state $\mathtt{s}_0$. (Many time we will assume that the initial state is deterministic and given by $\mathtt{s}_0$.) A control policy $\pi \in \Pi_{HS}$, together with the transition law $P = \{p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s},\mathtt{a})\}$ and the initial state distribution $p_0 = (p_0(\mathtt{s}),\ \mathtt{s} \in \mathcal{S}_0)$, induces a probability distribution over any finite state-action sequence $\mathtt{h}_{\mathtt{T}} = (\mathtt{s}_0, \mathtt{a}_0, \ldots, \mathtt{s}_{\mathtt{T}-1}, \mathtt{a}_{\mathtt{T}-1}, \mathtt{s}_{\mathtt{T}})$, given by

$$\Pr(\mathtt{h}_{\mathtt{T}}) = p_0(\mathtt{s}_0) \prod_{t=0}^{\mathtt{T}-1} p_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}+1}|\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) \pi_{\mathtt{t}}(\mathtt{a}_{\mathtt{t}}|\mathtt{h}_{\mathtt{t}}),$$

where $\mathtt{h}_{\mathtt{t}} = (\mathtt{s}_0, \mathtt{a}_0, \ldots, \mathtt{s}_{\mathtt{T}-1}, \mathtt{a}_{\mathtt{T}-1}, \mathtt{s}_{\mathtt{t}})$. To see this, observe the recursive relation:

$$\begin{aligned}\Pr(\mathtt{h}_{\mathtt{t}+1}) &= \Pr(\mathtt{h}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}, \mathtt{s}_{\mathtt{t}+1}) = \Pr(\mathtt{s}_{\mathtt{t}+1}|\mathtt{h}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) \Pr(\mathtt{a}_{\mathtt{t}}|\mathtt{h}_{\mathtt{t}}) \Pr(\mathtt{h}_{\mathtt{t}}) \\ &= p_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}+1}|\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) \pi_{\mathtt{t}}(\mathtt{a}_{\mathtt{t}}|\mathtt{h}_{\mathtt{t}}) \Pr(\mathtt{h}_{\mathtt{t}}).\end{aligned}$$

In the last step we used the conditional Markov property of the controlled chain: $\Pr(\mathtt{s}_{\mathtt{t}+1}|\mathtt{h}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) = p_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}+1}|\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}})$, and the definition of the control policy $\pi_{\mathtt{t}}$. The required formula follows by recursion.

Therefore, the state-action sequence $\mathtt{h}_\infty = (\mathtt{s}_k, \mathtt{a}_k)_{k \geq 0}$ can now be considered a stochastic process. We denote the probability law of this stochastic process by $\Pr^{\pi,p_0}(\cdot)$. The corresponding expectation operator is denoted by $\mathbb{E}^{\pi,p_0}(\cdot)$. When the initial state $\mathtt{s}_0$ is deterministic (i.e., $p_0(\mathtt{s})$ is concentrated on a single state $\mathtt{s}$), we may simply write $\Pr^{\pi,s}(\cdot)$ or $\Pr^{\pi}(\cdot|\mathtt{s}_0 = \mathtt{s})$.

Under a Markov control policy, the state sequence $(\mathtt{s}_{\mathtt{t}})_{\mathtt{t} \geq 0}$ becomes a *Markov chain*, with transition probabilities:

$$\Pr(\mathtt{s}_{\mathtt{t}+1} = \mathtt{s}'|\mathtt{s}_{\mathtt{t}} = \mathtt{s}) = \sum\nolimits_{\mathtt{a} \in \mathcal{A}_{\mathtt{t}}} p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s}, \mathtt{a}) \pi_{\mathtt{t}}(\mathtt{a}|\mathtt{s}).$$

This follows since:

$$\begin{aligned}\Pr(\mathtt{s}_{\mathtt{t}+1} = \mathtt{s}'|\mathtt{s}_{\mathtt{t}} = \mathtt{s}) &= \sum\nolimits_{\mathtt{a} \in \mathcal{A}_{\mathtt{t}}} \Pr(\mathtt{s}_{\mathtt{t}+1} = \mathtt{s}', \mathtt{a}|\mathtt{s}_{\mathtt{t}} = \mathtt{s}) \\ &= \sum\nolimits_{\mathtt{a} \in \mathcal{A}_{\mathtt{t}}} \Pr(\mathtt{s}_{\mathtt{t}+1} = \mathtt{s}'|\mathtt{s}_{\mathtt{t}} = \mathtt{s}, \mathtt{a}) \Pr(\mathtt{a}|\mathtt{s}_{\mathtt{t}} = \mathtt{s}) \\ &= \sum\nolimits_{\mathtt{a} \in \mathcal{A}_{\mathtt{t}}} p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s}, \mathtt{a}) \pi_{\mathtt{t}}(\mathtt{a}|\mathtt{s})\end{aligned}$$

If the controlled Markov chain is stationary (time-invariant) and the control policy is stationary, then the induced Markov chain is stationary as well.

**Remark 4.1.** *For most non-learning optimization problems, Markov policies suffice to achieve the optimum.*

**Remark 4.2.** *Implicit in these definitions of control policies is the assumption that the current state* $\mathtt{s_t}$ *can be fully observed before the action* $\mathtt{a_t}$ *is chosen . If this is not the case we need to consider the problem of a Partially Observed MDP (POMDP), which is more involved and will be discussed in Chapter 16.*

## 4.2 Performance Criteria

### 4.2.1 Finite Horizon Return

Consider the finite-horizon return, with a fixed time horizon $\mathtt{T}$. As in the deterministic case, we are given a running reward function $\mathtt{r_t} = \{\mathtt{r_t(s,a)} : \mathtt{s} \in \mathcal{S}_\mathtt{t}, \mathtt{a} \in \mathcal{A}_\mathtt{t}\}$ for $0 \le \mathtt{t} \le \mathtt{T}-1$, and a terminal reward function $\mathtt{r_T} = \{\mathtt{r_T(s)} : \mathtt{s} \in \mathcal{S}_\mathtt{T}\}$. The obtained reward is $R_\mathtt{t} = \mathtt{r_t(s_t, a_t)}$ at times $\mathtt{t} \le \mathtt{T}-1$, and $R_\mathtt{T} = \mathtt{r_T(s_T)}$ at the last stage. (Note that $\mathtt{s_t, a_t}$ and $\mathtt{s_T}$ are random variables that depend both on the policy $\pi$ and the stochastic transitions.) Our general goal is to maximize the cumulative return:

$$\sum_{t=0}^{\mathtt{T}} R_\mathtt{t} = \sum_{t=0}^{\mathtt{T}-1} \mathtt{r_t(s_t, a_t)} + \mathtt{r_T(s_T)}.$$

However, since the system is stochastic, the cumulative return will generally be a random variable, and we need to specify in which sense to maximize it. A natural first option is to consider the expected value of the return. That is, define:

$$\mathcal{V}_\mathtt{T}^\pi(\mathtt{s}) = \mathbb{E}^\pi(\sum_{t=0}^{\mathtt{T}} R_\mathtt{t} | \mathtt{s_0 = s}) \equiv \mathbb{E}^{\pi,s}(\sum_{t=0}^{\mathtt{T}} R_\mathtt{t}).$$

Here $\pi$ is the control policy as defined above, and $\mathtt{s}$ denotes the initial state. Hence, $\mathcal{V}_\mathtt{T}^\pi(\mathtt{s})$ is the expected cumulative return under the control policy $\pi$. Our goal is to find an optimal control policy that maximizes $\mathcal{V}_\mathtt{T}^\pi(\mathtt{s})$.

**Remarks:**

1. Dependence on the next state: In some problems, the obtained reward may depend on the next state as well: $R_\mathtt{t} = \tilde{\mathtt{r}}_\mathtt{t}(\mathtt{s_t, a_t, s_{t+1}})$. For control purposes, when we only consider the expected value of the reward, we can reduce this reward function to the usual one by defining

$$\mathtt{r_t(s,a)} \overset{\Delta}{=} \mathbb{E}(R_\mathtt{t} | \mathtt{s_t = s, a_t = a}) \equiv \sum\nolimits_{\mathtt{s}' \in S} p(\mathtt{s}'|\mathtt{s,a}) \tilde{r}_\mathtt{t}(s, a, \mathtt{s}')$$

2. Random rewards: The reward $R_\texttt{t}$ may also be random, namely a random variable whose distribution depends on $(\texttt{s}_\texttt{t}, \texttt{a}_\texttt{t})$. This can also be reduced to our standard model for planning purposes by looking at the expected value of $R_\texttt{t}$, namely
$$\texttt{r}_\texttt{t}(\texttt{s},\texttt{a}) = \mathbb{E}(R_\texttt{t}|\texttt{s}_\texttt{t} = \texttt{s}, \texttt{a}_\texttt{t} = \texttt{a}).$$

3. Risk-sensitive criteria: The expected cumulative return is by far the most common goal for planning. However, it is not the only one possible. For example, one may consider the following risk-sensitive return function:
$$\mathcal{V}^\pi_{\texttt{T},\lambda}(\texttt{s}) = \frac{1}{\lambda}\log \mathbb{E}^{\pi,s}(\exp(\lambda \sum_{t=0}^{\texttt{T}} R_\texttt{t})).$$
For $\lambda > 0$, the exponent gives higher weight to high rewards, and the opposite for $\lambda < 0$.

In the case that the rewards are stochastic, but have a discrete support, we can construct an equivalent MDP in which all the rewards are deterministic and has the same distribution of rewards. This implies that the important challenge is the stochastic state transition function, and the rewards can be assumed to be deterministic. Formally, given a trajectory we define a *rewards trajectory* as the sub-trajectory that includes only the rewards.

**Theorem 4.1.** *Given an MDP $M(\mathcal{S}, \mathcal{A}, p, \texttt{r}, \texttt{s}_0)$, where the rewards are stochastic, with support $\mathcal{K} = \{1, \dots, k\}$, there is an MDP $M'(\mathcal{S}\times\mathcal{K}, \mathcal{A}, p', \texttt{r}', \texttt{s}'_0)$, such that and a mapping of policies $\pi$ of $M$ to $\pi'$ policies of $M'$, such that: running $\pi$ in $M$ for horizon $\texttt{T}$ generates reward trajectory $R = (R_0, \dots, R_\texttt{T})$ and running $\pi'$ in $M'$ for horizon $\texttt{T}+1$ generates reward trajectory $R = (R_1, \dots, R_{\texttt{T}+1})$, then the distribution of $R$ and $R'$ are identical.*

*Proof.* The basic idea is to encode the rewards in the states of $M'$ which are $\mathcal{S}\times\mathcal{K} = \mathcal{S}'$. For each $(\texttt{s}, i) \in \mathcal{S}'$ and action $\texttt{a} \in \mathcal{A}$ we have $p'_t((\texttt{s}', j)|(\texttt{s}, i), \texttt{a}) = p_t(\texttt{s}'|\texttt{s},\texttt{a})\Pr[R_t(\texttt{s},\texttt{a}) = j]$, and $p'_\texttt{T}((\texttt{s}', j)|(\texttt{s}, i)) = \mathbb{I}(\texttt{s}' = \texttt{s})\Pr[R_\texttt{T}(\texttt{s}) = j]$. The reward is $\texttt{r}'_t((\texttt{s}, i), \texttt{a}) = i$. The initial state is $\texttt{s}'_0 = (\texttt{s}_0, 0)$.

For any policy $\pi(\texttt{a}|\texttt{s})$ in $M$ we have a policy $\pi'$ is $M'$ where $\pi'(\texttt{a}|(\texttt{s}, i)) = \pi(\texttt{a}|\texttt{s})$.

We map trajectories of $M$ to trajectories of $M'$ which have identical probabilities. A trajectory $(\texttt{s}_0, \texttt{a}_0, R_0, \texttt{s}_1, \texttt{a}_1, R_1, \texttt{s}_2 \dots, R_\texttt{T})$ is map to $((\texttt{s}_0, 0), \texttt{a}_0, 0, (\texttt{s}_1, R_0), \texttt{a}_1, R_0, (\texttt{s}_2, R_1) \dots, R_{\texttt{T}+1})$. Let $R$ and $R'$ be the respective reward trajectories. Clearly, the two trajectory have identical probabilities. This implies that the rewards trajectories $R$ and $R'$ have are identical probabilities (up to a shift of one in the index). □

Theorem 4.1 requires the number of rewards be bounded, and guarantees that the reward distribution be identical. In the case that the rewards are continuous, we can have a similar guarantee for linear return functions.

**Theorem 4.2.** *Given an MDP $M(\mathcal{S}, \mathcal{A}, p, \mathtt{r}, \mathtt{s}_0)$, where the rewards are stochastic, with support $[0, 1]$, there is an MDP $M'(\mathcal{S}, \mathcal{A}, p, \mathtt{r}', \mathtt{s}_0)$, where the rewards are stochastic, with support $\{0, 1\}$, such that for any policy $\pi \in \Pi_{MS}$ the distribution of the expected rewards trajectory is identical.*

*Proof.* We simply change the reward of $(\mathtt{s}, \mathtt{a})$ to be $\{0, 1\}$ by changing them to be a Bernoulli random variables with a parameter $\mathtt{r}_t(\mathtt{s}, \mathtt{a})$, i.e., $\Pr[\mathtt{R}_\mathtt{t}(\mathtt{s}, \mathtt{a}) = 1] = \mathtt{r}_\mathtt{t}(\mathtt{s}, \mathtt{a})$ and $\Pr[\mathtt{R}_\mathtt{t}(\mathtt{s}, \mathtt{a}) = 0] = 1 - \mathtt{r}_\mathtt{t}(\mathtt{s}, \mathtt{a})$. Clearly, the expected value of the rewards is identical. Further, since $\pi \in \Pi_{MS}$, it depends only of $\mathtt{s}$ and $\mathtt{t}$, which implies that the behavior (states and actions) will be identical in $M$ and $M'$. □

We have also established the following corollary.

**Corollary 4.3.** *Given an MDP $M(\mathcal{S}, \mathcal{A}, p, \mathtt{r}, \mathtt{s}_0)$, where the rewards are stochastic, with support $[0, 1]$, there is an MDP $M'(\mathcal{S} \times \{0, 1\}, \mathcal{A}, p', \mathtt{r}', \mathtt{s}'_0)$, and a mapping of policies $\pi \in \Pi_{MS}$ of $M$ to $\pi' \in \Pi_{MD}$ policies of $M'$, such that $\mathcal{V}^{\pi,M}_{\mathtt{T}}(\mathtt{s}_0) = \mathcal{V}^{\pi',M'}_{\mathtt{T}+1}(\mathtt{s}'_0)$*

### 4.2.2 Infinite Horizon Problems

We next consider planning problems that extend to an infinite time horizon, $\mathtt{t} = 0, 1, 2, \ldots$. Such planning problems arise when the system in question is expected to operate for a long time, or a large number of steps, possibly with no specific "closing" time. Infinite horizon problems are most often defined for stationary problems. In that case, they enjoy the important advantage that optimal policies can be found among the class of stationary policies. We will restrict attention here to stationary models. As before, we have the running reward function $\mathtt{r}(\mathtt{s}, \mathtt{a})$, which extends to all $\mathtt{t} \geq 0$. The expected reward obtained at stage $\mathtt{t}$ is $\mathbb{E}[R_\mathtt{t}] = \mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})$.

**Discounted return:** The most common performance criterion for infinite horizon problems is the expected discounted return:

$$\mathcal{V}^{\pi}_{\gamma}(\mathtt{s}) = \mathbb{E}^{\pi}(\sum_{t=0}^{\infty} \gamma^t \mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) | \mathtt{s}_0 = \mathtt{s}) \equiv \mathbb{E}^{\pi,s}(\sum_{t=0}^{\infty} \gamma^t \mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})) ,$$

where $0 < \gamma < 1$ is the discount factor. Mathematically, the discount factor ensures convergence of the sum (whenever the reward sequence is bounded). This make the problem "well behaved", and relatively easy to analyze. The discounted return is discussed in Chapter 5.

**Average return:** Here we are interested to maximize the long-term average return. The most common definition of the long-term average return is

$$\mathcal{V}^{\pi}_{av}(\mathtt{s}) = \liminf_{\mathtt{T}\to\infty} \mathbb{E}^{\pi,s}\left(\frac{1}{\mathtt{T}}\sum_{t=0}^{\mathtt{T}-1} \mathtt{r}(\mathtt{s_t}, \mathtt{a_t})\right)$$

The theory of average-return planning problems is more involved, and relies to a larger extent on the theory of Markov chains (see Chapter 3). The average return is discussed in Chapter 7.

### 4.2.3 Stochastic Shortest-Path Problems

In an important class of planning problems, the time horizon is not set beforehand, but rather the problem continues until a certain event occurs. This event can be defined as reaching some goal state. Let $\mathcal{S}_G \subset \mathcal{S}$ define the set of *goal states*. Define

$$\tau = \inf\{t \geq 0 : \mathtt{s_t} \in \mathcal{S}_G\}$$

as the first time in which a goal state is reached. The total expected return for this problem is defined as:

$$\mathcal{V}^{\pi}_{ssp}(\mathtt{s}) = \mathbb{E}^{\pi,\mathtt{s}}\left(\sum_{\mathtt{t}=0}^{\tau-1} \mathtt{r}(\mathtt{s_t}, \mathtt{a_t}) + \mathtt{r}_G(\mathtt{s}_\tau)\right)$$

Here $\mathtt{r}_G(\mathtt{s}),\ \mathtt{s} \in \mathcal{S}_G$ specified the reward at goal states. Note that the length of the run $\tau$ is a random variable.

Stochastic shortest path includes, naturally, the finite horizon case. This can be shown by creating a leveled MDP where at each time step we move to the next level and terminate at level $\mathtt{T}$. Specifically, we define a new state space $\mathcal{S}' = \mathcal{S} \times \mathbb{T}$, transition function $p((\mathtt{s}', i+1)|\mathtt{s}, \mathtt{a}) = p(\mathtt{s}'|\mathtt{s}, \mathtt{a})$, and goal states $\mathcal{S}_G = \{(\mathtt{s}, \mathtt{T}) : \mathtt{s} \in \mathcal{S}\}$.

Stochastic shortest path includes also the discounted infinite horizon. To see that, add a new goal state, and from each state with probability $1-\gamma$ jump to the goal state and terminate. The expected return of a policy would be the same in both models. Specifically, we add a state $\mathtt{s}_G$, such that $p(\mathtt{s}_G|\mathtt{s}, \mathtt{a}) = 1-\gamma$, for any state $\mathtt{s} \in \mathcal{S}$ and action $\mathtt{a} \in \mathcal{A}$ and $p(\mathtt{s}'|\mathtt{s}, \mathtt{a}) = \gamma p(\mathtt{s}'|\mathtt{s}, \mathtt{a})$. The probability that we do not terminate by time $\mathtt{t}$ is exactly $\gamma^{\mathtt{t}}$. Therefore, the expected return is $\sum_{t=1}^{\infty} \gamma^t \mathtt{r}(\mathtt{s}_t, \mathtt{a}_t)$ which is identical to the discounted return.

This class of problems provides a natural extension of the standard shortest-path problem to stochastic settings. Some conditions on the system dynamics and reward function must be imposed for the problem to be well posed (e.g., that a goal state may be reached with probability one). Such problems are known as stochastic shortest path problems, or also episodic MDP planning problems. See more in Chapter 6.

## 4.3 Sufficiency of Markov Policies

In all the performance criteria defined above, the criterion is composed of sums of terms of the form $\mathbb{E}(\mathtt{r_t}(\mathtt{s_t}, \mathtt{a_t}))$. It follows that if two control policies induce the same marginal probability distributions $q_t(\mathtt{s_t}, \mathtt{a_t})$ over the state-action pairs $(\mathtt{s_t}, \mathtt{a_t})$ for all $\mathtt{t} \geq 0$, they will have the same performance for any linear return function.

Using this observation, the next claim implies that it is enough to consider the set of (stochastic) Markov policies in the above planning problems.

**Proposition 4.4.** *Let $\pi \in \Pi_{HS}$ be a general (history-dependent, stochastic) control policy. Let*

$$p_{\mathtt{t}}^{\pi,\mathtt{s}_0}(\mathtt{s}, \mathtt{a}) = P^{\pi,\mathtt{s}_0}(\mathtt{s_t} = \mathtt{s}, \mathtt{a_t} = \mathtt{a}), \qquad (\mathtt{s}, \mathtt{a}) \in \mathcal{S}_{\mathtt{t}} \times \mathcal{A}_{\mathtt{t}}$$

*Denote the marginal distributions induced by $q_t(\mathtt{s_t}, \mathtt{a_t})$ on the state-action pairs $(\mathtt{s_t}, \mathtt{a_t})$, for all $\mathtt{t} \geq 0$. Then there exists a stochastic Markov policy $\tilde{\pi} \in \Pi_{MS}$ that induces the same marginal probabilities (for all initial states $\mathtt{s}_0$).*

In Chapter 2 we showed for Deterministic Decision Process for the finite horizon that there is an optimal deterministic policy. The proof that every stochastic history dependent strategy has an equivalent stochastic Markovian policy (Theorem 2.1) showed how to generate the same state-action distribution, and applies to other setting as well. The proof that every stochastic Markovian policy has a deterministic Markovian policy (Theorem 2.2) depended on the finite horizon, but it is easy to extend it to any linear return function as well.

## 4.4 Finite-Horizon Dynamic Programming

Recall that we consider the expected total reward criterion, which we denote as

$$\mathcal{V}^{\pi}(\mathtt{s}_0) = \mathbb{E}^{\pi,\mathtt{s}_0}\left(\sum_{\mathtt{t}=0}^{\mathtt{T}-1} \mathtt{r_t}(\mathtt{s_t}, \mathtt{a_t}) + \mathtt{r_T}(\mathtt{s_T})\right) ,$$

where $\pi$ is the control policy used, and $\mathtt{s}_0$ is a given initial state. We wish to maximize the expected return $\mathcal{V}^{\pi}(\mathtt{s}_0)$ over all control policies, and find an optimal policy $\pi^*$ that achieves the maximal expected return $\mathcal{V}^*(\mathtt{s}_0)$ for all initial states $\mathtt{s}_0$. Thus,

$$\mathcal{V}_{\mathtt{T}}^*(\mathtt{s}_0) \triangleq \mathcal{V}_{\mathtt{T}}^{\pi^*}(\mathtt{s}_0) = \max_{\pi \in \Pi_{HS}} \mathcal{V}_{\mathtt{T}}^{\pi}(\mathtt{s}_0)$$

### 4.4.1 The Principle of Optimality

The celebrated principle of optimality (stated by Bellman) applies to a large class of multi-stage optimization problems, and is at the heart of Dynamic Programming. As a general principle, it states that:

**The tail of an optimal policy is optimal for the "tail" problem.**

This principle is not an actual claim, but rather a guiding principle that can be applied in different ways to each problem. For example, considering our finite-horizon problem, let $\pi^* = (\pi_0, \ldots, \pi_{\mathtt{T}-1})$ denote an optimal Markov policy. Take any state $\mathtt{s_t} = \mathtt{s}'$ which has a positive probability to be reached under $\pi^*$, namely $P^{\pi^*,\mathtt{s}_0}(\mathtt{s_t} = \mathtt{s}') > 0$. Then the tail policy $\pi^*_{t:T} = (\pi_\mathtt{t}, \ldots, \pi_{\mathtt{T}-1})$ is optimal for the "tail" criterion $\mathcal{V}^\pi_{t:T}(\mathtt{s}') = \mathbb{E}^\pi\left(\sum_{k=t}^{\mathtt{T}} R_k | \mathtt{s_t} = \mathtt{s}'\right)$.

Note that the reverse is not true. The prefix of the optimal policy is not optimal for the "prefix" problem. When we plan for a long horizon, we might start with non-greedy actions, so we can improve our return in later time steps. Specifically, the first action taken does not have to be the optimal action for horizon $\mathtt{T} = 1$, for which the greedy action is optimal.

### 4.4.2 Dynamic Programming for Policy Evaluation

As a "warmup", let us evaluate the reward of a given policy. Let $\pi = (\pi_0, \ldots, \pi_{\mathtt{T}-1})$ be a given Markov policy. Define the following reward-to-go function, or value function:

$$V_k^\pi(\mathtt{s}) = \mathbb{E}^\pi\left(\sum\nolimits_{\mathtt{t}=k}^{\mathtt{T}} R_\mathtt{t} | \mathtt{s}_k = \mathtt{s}\right)$$

Observe that $V_0^\pi(\mathtt{s}_0) = \mathcal{V}^\pi(\mathtt{s}_0)$.

**Lemma 4.5 (Value Iteration).** *$V_k^\pi(\mathtt{s})$ may be computed by the backward recursion:*

$$V_k^\pi(\mathtt{s}) = \left\{\mathtt{r}_k(\mathtt{s},\mathtt{a}) + \sum\nolimits_{\mathtt{s}'\in\mathcal{S}_{k+1}} p_k(\mathtt{s}'|\mathtt{s},\mathtt{a})\, V_{k+1}^\pi(\mathtt{s}')\right\}_{a=\pi_k(\mathtt{s})}, \quad \mathtt{s} \in \mathcal{S}_k$$

*for $k = \mathtt{T}-1, \ldots, 0$, starting with $V_\mathtt{T}^\pi(\mathtt{s}) = \mathtt{r}_\mathtt{T}(\mathtt{s})$.*

*Proof.* Observe that:

$$\begin{aligned}
V_k^\pi(\mathtt{s}) &= \mathbb{E}^\pi\left(R_k + \sum\nolimits_{\mathtt{t}=k+1}^{\mathtt{T}} R_\mathtt{t} |\ \mathtt{s}_k = \mathtt{s}, \mathtt{a}_k = \pi_k(\mathtt{s})\right) \\
&= \mathbb{E}^\pi\left(\mathbb{E}^\pi\left(R_k + \sum\nolimits_{\mathtt{t}=k+1}^{\mathtt{T}} R_\mathtt{t} |\ \mathtt{s}_k = \mathtt{s}, \mathtt{a}_k = \pi_k(\mathtt{s}), \mathtt{s}_{k+1}\right) | \mathtt{s}_k = \mathtt{s}, \mathtt{a}_k = \pi_k(\mathtt{s})\right) \\
&= \mathbb{E}^\pi\left(\mathtt{r}_k(\mathtt{s}_k, \mathtt{a}_k) + V_{k+1}^\pi(\mathtt{s}_{k+1}) | \mathtt{s}_k = \mathtt{s}, \mathtt{a}_k = \pi_k(\mathtt{s})\right) \\
&= \mathtt{r}_k(s, \pi_k(\mathtt{s})) + \sum\nolimits_{\mathtt{s}'\in\mathcal{S}_{k+1}} p_k(\mathtt{s}'|s, \pi_k(\mathtt{s}))\, V_{k+1}^\pi(\mathtt{s}')
\end{aligned}$$

The first identity is simply writing the value function explicitly, starting at state $\mathtt{s}$ at time $k$ and using action $\mathtt{a} = \pi_k(\mathtt{s})$. We split the sum to $R_k$, immediate reward, and the sum of other rewards. The second identity uses the law of total probability, we are conditioning on state $\mathtt{s}_{k+1}$, and taking the expectation over it. The third identity observes that the expected value of the sum is actually the value function at $\mathtt{s}_{k+1}$. The last identity writes the expectation over $\mathtt{s}_{k+1}$ explicitly. $\square$

**Remarks:**

- Note that $\sum_{\mathtt{s}'\in\mathcal{S}_{k+1}} p_k(\mathtt{s}'|\mathtt{s},\mathtt{a})\, V^\pi_{k+1}(\mathtt{s}') = \mathbb{E}^\pi(V^\pi_{k+1}(\mathtt{s}_{k+1})|\mathtt{s}_k = \mathtt{s}, \mathtt{a}_k = \mathtt{a})$.
- For the more general reward function $\tilde{\mathtt{r}}_\mathtt{t}(\mathtt{s},\mathtt{a},\mathtt{s}')$, the recursion takes the form

$$V^\pi_k(\mathtt{s}) = \left\{\sum\nolimits_{\mathtt{s}'\in\mathcal{S}_{k+1}} p_k(\mathtt{s}'|\mathtt{s},\mathtt{a})[\tilde{\mathtt{r}}_k(\mathtt{s},\mathtt{a},\mathtt{s}') + V^\pi_{k+1}(\mathtt{s}')]\right\}_{\mathtt{a}=\pi_k(\mathtt{s})} .$$

  A similar observation applies to the Dynamic Programming equations in the next section.

### 4.4.3 Dynamic Programming for Policy Optimization

We next define the optimal value function at each time $k \geq 0$ :

$$V^*_k(\mathtt{s}) = \max_{\pi^k} \mathbb{E}^{\pi^k}\left(\sum\nolimits_{\mathtt{t}=k}^{\mathtt{T}} R_\mathtt{t}|\mathtt{s}_k = \mathtt{s}\right), \quad \mathtt{s} \in \mathcal{S}_k ,$$

where the maximum is taken over "tail" policies $\pi^k = (\pi_k, \dots, \pi_{\mathtt{T}-1})$ that start from time $k$. Note that $\pi^k$ is allowed to be a general policy, i.e., history-dependent and stochastic. Obviously, $V_0(\mathtt{s}_0) = \mathcal{V}^*(\mathtt{s}_0)$.

**Theorem 4.6** (**Finite-horizon Dynamic Programming**). *The following holds:*

1. *Backward recursion: Set* $V_\mathtt{T}(\mathtt{s}) = \mathtt{r}_\mathtt{T}(\mathtt{s})$ *for* $\mathtt{s} \in \mathcal{S}_\mathtt{T}$.
   *For* $k = \mathtt{T}-1, \dots, 0$, $V_k(\mathtt{s})$ *compute using the following recursion:*

$$V_k(\mathtt{s}) = \max_{\mathtt{a}\in\mathcal{A}_k}\left\{\mathtt{r}_k(\mathtt{s},\mathtt{a}) + \sum\nolimits_{\mathtt{s}'\in\mathcal{S}_{k+1}} p_k(\mathtt{s}'|\mathtt{s},\mathtt{a})\, V_{k+1}(\mathtt{s}')\right\}, \quad \mathtt{s} \in \mathcal{S}_k.$$

   *We have that* $V_k(s) = V^*_k(s)$.

2. *Optimal policy: Any Markov policy $\pi^*$ that satisfies, for $\mathtt{t} = 0, \ldots, \mathtt{T}-1$,*

$$\pi^*_{\mathtt{t}}(\mathtt{s}) \in \arg\max_{\mathtt{a}\in\mathcal{A}_{\mathtt{t}}} \left\{ \mathtt{r}_{\mathtt{t}}(\mathtt{s},\mathtt{a}) + \sum\nolimits_{\mathtt{s}'\in\mathcal{S}_{\mathtt{t}+1}} p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s},\mathtt{a})\, V_{\mathtt{t}+1}(\mathtt{s}') \right\}, \quad \mathtt{s}\in\mathcal{S}_{\mathtt{t}},$$

*is an optimal control policy. Furthermore, $\pi^*$ maximizes $\mathcal{V}^{\pi}(\mathtt{s}_0)$ simultaneously for every initial state $\mathtt{s}_0 \in \mathcal{S}_0$.*

Note that Theorem 4.6 specifies an optimal control policy which is a deterministic Markov policy.

*Proof.* **Part (i):**
We use induction to show that the stated backward recursion indeed yields the optimal value function $V^*_{\mathtt{t}}$. The idea is simple, but some care is needed with the notation since we consider general policies, and not just Markov policies.

For the base of the induction we start with $\mathtt{t} = \mathtt{T}$. The equality $V_{\mathtt{T}}(\mathtt{s}) = \mathtt{r}_{\mathtt{T}}(\mathtt{s})$ follows directly from the definition of $V_{\mathtt{T}}$. Clearly this is also the optimal value function $V^*_{\mathtt{T}}$.

We proceed by backward induction. Suppose that $V_{k+1}(\mathtt{s})$ is the optimal value function for time $k+1$, i.e., $V_{k+1}(\mathtt{s}) = V^*_{k+1}$ . We need to show that $V_k(\mathtt{s}) = V^*_k(\mathtt{s})$ and we do it by showing that $V^*_k(\mathtt{s}) = W_k(\mathtt{s})$, where

$$W_k(\mathtt{s}) \triangleq \max_{\mathtt{a}\in\mathcal{A}_k} \left\{ \mathtt{r}_k(\mathtt{s},\mathtt{a}) + \sum\nolimits_{\mathtt{s}'\in\mathcal{S}_{k+1}} p_k(\mathtt{s}'|\mathtt{s},\mathtt{a})\, V_{k+1}(\mathtt{s}') \right\}.$$

We will first establish that $V^*_k(\mathtt{s}) \geq W_k(\mathtt{s})$, and then that $V^*_k(\mathtt{s}) \leq W_k(\mathtt{s})$.

(a) We first show that $V^*_k(\mathtt{s}) \geq W_k(\mathtt{s})$. For that purpose, it is enough to find a policy $\pi^k$ so that $V^{\pi^k}_k(\mathtt{s}) = W_k(\mathtt{s})$, since $V^*_k(s) \geq V^{\pi}_k(s)$ for any strategy $\pi$.

Fix $\mathtt{s} \in \mathcal{S}_k$, and define $\pi^k$ as follows: Choose $\mathtt{a}_k = \bar{\mathtt{a}}$, where

$$\bar{\mathtt{a}} \in \arg\max_{\mathtt{a}\in\mathcal{A}_k} \left\{ \mathtt{r}_k(\mathtt{s},\mathtt{a}) + \sum\nolimits_{\mathtt{s}'\in\mathcal{S}_{k+1}} p_k(\mathtt{s}'|\mathtt{s},\mathtt{a})\, V_{k+1}(\mathtt{s}') \right\},$$

and then, after observing $\mathtt{s}_{k+1} = \mathtt{s}'$, proceed with the optimal tail policy $\pi^{k+1}(\mathtt{s}')$ that obtains $V^{\pi^{k+1}(\mathtt{s}')}_{k+1}(\mathtt{s}') = V_{k+1}(\mathtt{s}')$. Proceeding similarly to the proof of Lemma 4.5 (value iteration for a fixed policy), we obtain:

$$V^{\pi^k}_k(\mathtt{s}) = \mathtt{r}_k(\mathtt{s},\bar{\mathtt{a}}) + \sum\nolimits_{\mathtt{s}'\in\mathcal{S}_{k+1}} p(\mathtt{s}'|\mathtt{s},\bar{\mathtt{a}})\, V^{\pi^{k+1}(\mathtt{s}')}_{k+1}(\mathtt{s}') \tag{4.1}$$

$$= \mathtt{r}_k(\mathtt{s},\bar{\mathtt{a}}) + \sum\nolimits_{\mathtt{s}'\in\mathcal{S}_{k+1}} p(\mathtt{s}'|\mathtt{s},\bar{\mathtt{a}})\, V_{k+1}(\mathtt{s}') = W_k(\mathtt{s}), \tag{4.2}$$

as was required.

(b) To establish $V_k^*(\mathtt{s}) \leq W_k(\mathtt{s})$, it is enough to show that $V_k^{\pi^k}(\mathtt{s}) \leq W_k(\mathtt{s})$ for any (general, randomized) "tail" policy $\pi^k$.

Fix $\mathtt{s} \in \mathcal{S}_k$. Consider then some tail policy $\pi^k = (\pi_k, \dots \pi_{\mathtt{T}-1})$. Note that this means that $\mathtt{a}_\mathtt{t} \sim \pi_\mathtt{t}(\mathtt{a}|\mathtt{h}_{k:t})$, where $\mathtt{h}_{k:t} = (\mathtt{s}_k, \mathtt{a}_k, \mathtt{s}_{k+1}, \mathtt{a}_{k+1}, \dots, \mathtt{s}_\mathtt{t})$. For each state-action pair $\mathtt{s} \in \mathcal{S}_k$ and $\mathtt{a} \in \mathcal{A}_k$, let $(\pi^k|\mathtt{s},\mathtt{a})$ denote the tail policy $\pi^{k+1}$ from time $k+1$ onwards which is obtained from $\pi^k$ given that $\mathtt{s}_k = \mathtt{s}$, $\mathtt{a}_k = \mathtt{a}$. As before, by value iteration for a fixed policy,

$$V_k^{\pi^k}(\mathtt{s}) = \sum\nolimits_{\mathtt{a}\in\mathcal{A}_k} \pi_k(\mathtt{a}|\mathtt{s}) \left\{ \mathtt{r}_k(\mathtt{s},\mathtt{a}) + \sum\nolimits_{\mathtt{s}'\in\mathcal{S}_{k+1}} p_k(\mathtt{s}'|\mathtt{s},\mathtt{a})\, V_{k+1}^{(\pi^k|\mathtt{s},\mathtt{a})}(\mathtt{s}') \right\}.$$

But since $V_{k+1}$ is optimal,

$$\begin{aligned} V_k^{\pi^k}(\mathtt{s}) &\leq \sum\nolimits_{\mathtt{a}\in\mathcal{A}_k} \pi_k(\mathtt{a}|\mathtt{s}) \left\{ \mathtt{r}_k(\mathtt{s},\mathtt{a}) + \sum\nolimits_{\mathtt{s}'\in\mathcal{S}_{k+1}} p_k(\mathtt{s}'|\mathtt{s},\mathtt{a})\, V_{k+1}(\mathtt{s}') \right\} \\ &\leq \max_{\mathtt{a}\in\mathcal{A}_k} \left\{ \mathtt{r}_k(\mathtt{s},\mathtt{a}) + \sum\nolimits_{\mathtt{s}'\in S_{k+1}} p_k(\mathtt{s}'|\mathtt{s},\mathtt{a})\, V_{k+1}(\mathtt{s}') \right\} = W_k(\mathtt{s}), \end{aligned}$$

which is the required inequality in (b).

**Part (ii)** The main point is to show that the it is sufficient that the optimal policy would be Markov (rather than history dependent) and deterministic (rather than stochastic).

We will only sketch the proof. Let $\pi^*$ be the (Markov) policy defined in part 2 of Theorem 4.6. Our goal is to show that the value function of $\pi^*$ coincides with that of the optima policy, which we showed is equal to $V_k$ that we computed. Once we show that, we prove that $\pi^*$ is optimal.

Consider the value iteration (Lemma 4.5). The updates for $V_k^{\pi^*}$ in the value iteration, given the action selection of $\pi^*$, are identical to those of $V_k$. This implies that $V_k^{\pi^*} = V_k$ (formally, by induction of $k$). Since $V_k$ is the optimal value function, it implies that $\pi^*$ is the optimal policy. □

### 4.4.4 The Q function

Let

$$Q_k^*(\mathtt{s},\mathtt{a}) \stackrel{\Delta}{=} \mathtt{r}_k(\mathtt{s},\mathtt{a}) + \sum\nolimits_{\mathtt{s}'\in S_k} p_k(\mathtt{s}'|\mathtt{s},\mathtt{a})\, V_k^*(\mathtt{s}').$$

This is known as the optimal state-action value function, or simply as the *Q-function.* $Q_k^*(\mathtt{s},\mathtt{a})$ is the expected return from stage $k$ onward, if we choose $\mathtt{a}_k = \mathtt{a}$ and then proceed optimally.

Theorem 4.6 can now be succinctly expressed as

$$V_k^*(\mathtt{s}) = \max_{\mathtt{a}\in\mathcal{A}_k} Q_k^*(\mathtt{s},\mathtt{a}),$$

and

$$\pi_k^*(\mathtt{s}) \in \arg\max_{\mathtt{a}\in\mathcal{A}_k} Q_k^*(\mathtt{s},\mathtt{a}).$$

The Q function provides the basis for the Q-learning algorithm, which is one of the basic Reinforcement Learning algorithms, and would be discussed in Chapter 9.

## 4.5 Linear Program

In this section we will extend the linear programming given in Section 2.5 from Deterministic Decision Processes to Markov Decision Processes. The derivation is almost identical and the difference is that replacing the deterministic next state $f_{\mathtt{t}}(\mathtt{s},\mathtt{a})$ by an expectation over $\mathtt{s}'$, such that each $\mathtt{s}'$ has a probability $p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s},\mathtt{a})$. For completeness we do it in a detailed way.

We will see that both the primal and dual program will play an important part in defining the optimal policy. We will fix an initial state $\mathtt{s}_0$ and compute the optimal policy for it.

We will start with the primal linear program, which will compute the optimal policy. For each time $\mathtt{t}$, state $\mathtt{s}$ and action $\mathtt{a}$ we will have a variable $x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) \in [0,1]$ that will indicate by probability that at time $\mathtt{t}$ we are at state $\mathtt{s}$ and perform action $\mathtt{a}$. For the terminal states $\mathtt{s}$ we will have a variable $x_{\mathtt{T}}(\mathtt{s}) \in [0,1]$ that will indicate whether the probability that we terminate at state $\mathtt{s}$.

Our main constraint will be a flow constraint, stating that the probability mass that leaves a state $\mathtt{s}$ at time $\mathtt{t}$ is bounded by the probability mass of reaching state $\mathtt{s}$ at time $\mathtt{t}-1$. Formally,

$$\sum_{\mathtt{a}} x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) \le \sum_{\mathtt{s}',\mathtt{a}'} x_{\mathtt{t}-1}(\mathtt{s}',\mathtt{a}') p_{\mathtt{t}-1}(\mathtt{s}|\mathtt{s}'\mathtt{a}').$$

and for terminal states simply

$$x_{\mathtt{T}}(\mathtt{s}) \le \sum_{\mathtt{s}',\mathtt{a}'} x_{\mathtt{T}-1}(\mathtt{s}',\mathtt{a}') p_{\mathtt{T}-1}(\mathtt{s}|\mathtt{s}',\mathtt{a}')$$

The return, which we would like to maximize, would be

$$\sum_{\mathtt{t},\mathtt{s},\mathtt{a}} \mathtt{r}_{\mathtt{t}}(\mathtt{s},\mathtt{a}) x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) + \sum_{\mathtt{s}} \mathtt{r}_{\mathtt{T}}(\mathtt{s}) x_{\mathtt{T}}(\mathtt{s})$$

The resulting linear program is the following.

$$\begin{aligned}
\max_{x_{\mathtt{t}}(\mathtt{s},\mathtt{a}),x_{\mathtt{T}}(\mathtt{s})} \quad & \sum_{\mathtt{t},\mathtt{s},\mathtt{a}} \mathtt{r}_{\mathtt{t}}(\mathtt{s},\mathtt{a})x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) + \sum_{\mathtt{s}} \mathtt{r}_{\mathtt{T}}(\mathtt{s})x_{\mathtt{T}}(\mathtt{s}) \\
& \text{such that} \\
& \sum_{\mathtt{a}} x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) \le \sum_{\mathtt{s}',\mathtt{a}'} x_{\mathtt{t}-1}(\mathtt{s}',\mathtt{a}')p_{\mathtt{t}-1}(\mathtt{s}|\mathtt{s}'\mathtt{a}'). && \forall \mathtt{s}\in\mathcal{S}_{\mathtt{t}}, \mathtt{t}\in\mathbb{T} \\
& x_{\mathtt{T}}(\mathtt{s}) \le \sum_{\mathtt{s}',\mathtt{a}'} x_{\mathtt{T}-1}(\mathtt{s}',\mathtt{a}')p_{\mathtt{T}-1}(\mathtt{s}|\mathtt{s}'\mathtt{a}') && \forall \mathtt{s}\in\mathcal{S}_{\mathtt{T}} \\
& x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) \ge 0 && \forall \mathtt{s}\in\mathcal{S}_{\mathtt{t}}, \mathtt{a}\in\mathcal{A}, \mathtt{t}\in\{0,\dots,\mathtt{T}-1\} \\
& \sum_{\mathtt{a}} x_0(\mathtt{s}_0,\mathtt{a}) = 1 \\
& x_0(\mathtt{s},\mathtt{a}) = 0, && \forall \mathtt{s}\in\mathcal{S}_0, \mathtt{s}\ne\mathtt{s}_0
\end{aligned}$$

At first sight it is surprising that we do not impose that $x_{\mathtt{t}}(\mathtt{s},\mathtt{a}) \le 1$, however this is implicit in the program. Let $\Phi(\mathtt{t}) = \sum_{\mathtt{s},\mathtt{a}} x_{\mathtt{t}}(\mathtt{s},\mathtt{a})$. From the initial conditions we have that $\Phi(0) = 1$. When we sum over the state the flow condition (first inequality) we have that $\Phi(\mathtt{t}) \le \Phi(\mathtt{t}-1)$. This implies that $\Phi(\mathtt{t}) \le 1$.

Given the primal linear program we can derive the dual linear program.

$$\begin{aligned}
\min_{z_{\mathtt{t}}(\mathtt{s})} \; & z_0(\mathtt{s}_0) \\
& \text{such that} \\
& z_{\mathtt{T}}(\mathtt{s}) = \mathtt{r}_{\mathtt{T}}(\mathtt{s}) && \forall \mathtt{s}\in\mathcal{S}_{\mathtt{t}} \\
& z_{\mathtt{t}}(\mathtt{s}) \ge \mathtt{r}_{\mathtt{t}}(\mathtt{s},\mathtt{a}) + \sum_{\mathtt{s}'} z_{\mathtt{t}+1}(\mathtt{s}')p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s},\mathtt{a}), && \forall \mathtt{s}\in\mathcal{S}_{\mathtt{t}}, \mathtt{a}\in\mathcal{A}, \mathtt{t}\in\mathbb{T}, \\
& .
\end{aligned}$$

One can identify the dual random variables $z_{\mathtt{t}}(\mathtt{s})$ with the optimal vale function $\mathcal{V}_{\mathtt{t}}(\mathtt{s})$. At the optimal solution of the dual linear program one can show that we have

$$z_{\mathtt{t}}(\mathtt{s}) = \max_{\mathtt{a}} \big\{\mathtt{r}_{\mathtt{t}}(\mathtt{s},\mathtt{a}) + \sum_{\mathtt{s}'} z_{\mathtt{t}+1}(\mathtt{s}')p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s},\mathtt{a})\big\}, \qquad \forall \mathtt{s}\in\mathcal{S}_{\mathtt{t}}, \mathtt{t}\in\mathbb{T},$$

which are the familiar Bellman optimality equations.

## 4.6 Summary

- The optimal value function can be computed by backward recursion. This recursive equation is known as the *dynamic programming equation*, *optimality equation*, or *Bellman's Equation.*
- Computation of the value function in this way is known as the *finite-horizon value iteration* algorithm.
- The value function is computed for all states at each stage.
- An optimal policy is easily derived from the optimal value.
- The optimization in each stage is performed in the action space. The total number of minimization operations needed is $\mathtt{T} \times |\mathcal{S}|$ - each over $|\mathcal{A}|$ choices. This replaces "brute force" optimization in policy space, with tremendous computational savings as the number of Markov policies is $|\mathcal{A}|^{\mathtt{T}\times|\mathcal{S}|}$.

# Chapter 5

# MDPs with Discounted Infinite Return

This chapter covers the basic theory and main solution methods for stationary MDPs over an infinite horizon, with the discounted return criterion. In this case, we will show that stationary policies are optimal.

The discounted return problem is the most "well behaved" among all infinite horizon problems (such as average return and stochastic shortest path), and the theory of it is relatively simple, both in the planning and the learning contexts. For that reason, as well as its usefulness, we will consider here the discounted problem and its solution in some detail.

## 5.1 Problem Statement

We consider a stationary (time-invariant) MDP, with a finite state space $\mathcal{S}$, finite action set $\mathcal{A}$, and transition kernel $P = (P(\mathtt{s}'|\mathtt{s},\mathtt{a}))$ over the infinite time horizon $\mathbb{T} = \{0, 1, 2, \ldots\}$.

Our goal is to maximize the expected discounted return, which is defined for each control policy $\pi$ and initial state $\mathtt{s}_0 = s$ as follows:

$$\begin{aligned}\mathcal{V}_\gamma^\pi(\mathtt{s}) &= \mathbb{E}^\pi(\sum_{\mathtt{t}=0}^{\infty} \gamma^{\mathtt{t}} \mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})|\mathtt{s}_0 = \mathtt{s}) \\ &\equiv \mathbb{E}^{\pi,\mathtt{s}}(\sum_{\mathtt{t}=0}^{\infty} \gamma^{\mathtt{t}} \mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}))\end{aligned}$$

where $\mathbb{E}^{\pi,\mathtt{s}}$ uses the distribution induced by policy $\pi$ starting at state $\mathtt{s}$. Here,

- $\mathtt{r} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the (running, or instantaneous) expected reward function, i.e., $\mathtt{r}(\mathtt{s}, \mathtt{a}) = \mathbb{E}[R|\mathtt{s}, \mathtt{a}]$.
- $\gamma \in (0, 1)$ is the discount factor.

We observe that $\gamma < 1$ ensures convergence of the infinite sum (since the rewards $\mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})$ are uniformly bounded). With $\gamma = 1$ we obtain the total return criterion, which is harder to handle due to possible divergence of the sum.

Let $\mathcal{V}^*_\gamma(\mathtt{s})$ denote the maximal expected value of the discounted return, over all (possibly history dependent and randomized) control policies, i.e.,

$$\mathcal{V}^*_\gamma(\mathtt{s}) = \sup_{\pi \in \Pi_{HS}} \mathcal{V}^\pi_\gamma(\mathtt{s}).$$

Our goal is to find an optimal control policy $\pi^*$ that attains that maximum (for all initial states), and compute the numeric value of the optimal return $\mathcal{V}^*_\gamma(\mathtt{s})$. As we shall see, for this problem there always exists an optimal policy which is a (deterministic) stationary policy.

**Note:** As usual, the discounted performance criterion can be defined in terms of cost:

$$\mathcal{C}^\pi_\gamma(\mathtt{s}) = \mathbb{E}^{\pi,\mathtt{s}}(\sum_{\mathtt{t}=0}^{\infty} \gamma^\mathtt{t} \mathtt{c}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})) ,$$

where $\mathtt{c}(\mathtt{s}, \mathtt{a})$ is the running cost function. Our goal is then to minimize the discounted cost $\mathcal{C}^\pi_\gamma(\mathtt{s})$.

## 5.2 The Fixed-Policy Value Function

We start the analysis by defining and computing the value function for a fixed stationary policy. This intermediate step is required for later analysis of our optimization problem, and also serves as a gentle introduction to the value iteration approach.

For a stationary policy $\pi : \mathcal{S} \to \mathcal{A}$, we define the value function $\mathcal{V}^\pi(\mathtt{s})$, $\mathtt{s} \in \mathcal{S}$ simply as the corresponding discounted return:

$$\mathcal{V}^\pi(\mathtt{s}) \triangleq \mathbb{E}^{\pi,\mathtt{s}} \left( \sum_{\mathtt{t}=0}^{\infty} \gamma^\mathtt{t} \mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) \right) = \mathcal{V}^\pi_\gamma(\mathtt{s}), \quad \mathtt{s} \in S$$

**Lemma 5.1.** *For $\pi \in \Pi_{SD}$, the value function $\mathcal{V}^\pi$ satisfies the following set of $|\mathcal{S}|$ linear equations:*

$$\mathcal{V}^\pi(\mathtt{s}) = \mathtt{r}(\mathtt{s}, \pi(\mathtt{s})) + \gamma \sum\nolimits_{\mathtt{s}' \in \mathcal{S}} p(\mathtt{s}'|\mathtt{s}, \pi(\mathtt{s})) \mathcal{V}^\pi(\mathtt{s}') , \quad \mathtt{s} \in \mathcal{S}. \tag{5.1}$$

*Proof.* We first note that

$$\begin{aligned}\mathcal{V}^{\pi}(\mathtt{s}) &\triangleq \mathbb{E}^{\pi}(\sum_{\mathtt{t}=0}^{\infty}\gamma^{\mathtt{t}}\mathtt{r}(\mathtt{s}_{\mathtt{t}},\mathtt{a}_{\mathtt{t}})|\mathtt{s}_0=\mathtt{s}) \\ &= \mathbb{E}^{\pi}(\sum_{\mathtt{t}=1}^{\infty}\gamma^{t-1}\mathtt{r}(\mathtt{s}_{\mathtt{t}},\mathtt{a}_{\mathtt{t}})|\mathtt{s}_1=\mathtt{s}),\end{aligned}$$

since both the model and the policy are stationary. Now,

$$\begin{aligned}\mathcal{V}^{\pi}(\mathtt{s}) &= \mathtt{r}(\mathtt{s},\pi(\mathtt{s})) + \mathbb{E}^{\pi}(\sum_{\mathtt{t}=1}^{\infty}\gamma^{\mathtt{t}}\mathtt{r}(\mathtt{s}_{\mathtt{t}},\pi(\mathtt{s}_{\mathtt{t}}))|\mathtt{s}_0=\mathtt{s}) \\ &= \mathtt{r}(\mathtt{s},\pi(\mathtt{s})) + \mathbb{E}^{\pi}\left[\mathbb{E}^{\pi}\left(\sum_{\mathtt{t}=1}^{\infty}\gamma^{\mathtt{t}}\mathtt{r}(\mathtt{s}_{\mathtt{t}},\pi(\mathtt{s}_{\mathtt{t}}))|\mathtt{s}_0=\mathtt{s},\mathtt{s}_1=\mathtt{s}'\right)\middle|\,\mathtt{s}_0=s\right] \\ &= \mathtt{r}(\mathtt{s},\pi(\mathtt{s})) + \sum_{\mathtt{s}'\in\mathcal{S}}p(\mathtt{s}'|\mathtt{s},\pi(\mathtt{s}))\mathbb{E}^{\pi}(\sum_{\mathtt{t}=1}^{\infty}\gamma^{\mathtt{t}}\mathtt{r}(\mathtt{s}_{\mathtt{t}},\pi(\mathtt{s}_{\mathtt{t}}))|\mathtt{s}_1=\mathtt{s}') \\ &= \mathtt{r}(\mathtt{s},\pi(\mathtt{s})) + \gamma\sum_{\mathtt{s}'\in\mathcal{S}}p(\mathtt{s}'|\mathtt{s},\pi(\mathtt{s}))\mathbb{E}^{\pi}(\sum_{\mathtt{t}=1}^{\infty}\gamma^{t-1}\mathtt{r}(\mathtt{s}_{\mathtt{t}},\mathtt{a}_{\mathtt{t}})|\mathtt{s}_1=\mathtt{s}') \\ &= \mathtt{r}(\mathtt{s},\pi(\mathtt{s})) + \gamma\sum_{\mathtt{s}'\in\mathcal{S}}p(\mathtt{s}'|\mathtt{s},\pi(\mathtt{s}))\mathcal{V}^{\pi}(\mathtt{s}').\end{aligned}$$

The first equality is by the definition of the value function. The second equality follows from the law of total expectation, conditioning $\mathtt{s}_1 = \mathtt{s}'$ and taking the expectation over it. By definition $\mathtt{a}_{\mathtt{t}} = \pi(\mathtt{s}_{\mathtt{t}})$. The third equality follows similarly to the finite-horizon case (Lemma 4.5, in Chapter 2.1). The fourth is simple algebra, taking a one multiple of the discount factor $\gamma$ outside. The last by the observation in the beginning of the proof. □

We can write the linear equations in (5.1) in vector form as follows. Define the column vector $r^{\pi} = (r^{\pi}(\mathtt{s}))_{\mathtt{s}\in\mathcal{S}}$ with components $r^{\pi}(\mathtt{s}) = \mathtt{r}(\mathtt{s},\pi(\mathtt{s}))$, and the transition matrix $P^{\pi}$ with components $P^{\pi}(\mathtt{s}'|\mathtt{s}) = p(\mathtt{s}'|\mathtt{s},\pi(\mathtt{s}))$. Finally, let $\mathcal{V}^{\pi}$ denote a column vector with components $\mathcal{V}^{\pi}(\mathtt{s})$. Then (5.1) is equivalent to the linear equation set

$$\mathcal{V}^{\pi} = r^{\pi} + \gamma P^{\pi}\mathcal{V}^{\pi} \tag{5.2}$$

**Lemma 5.2.** *The set of linear equations* (5.1) *or* (5.2)*, with $\mathcal{V}^{\pi}$ as variables, has a unique solution $\mathcal{V}^{\pi}$, which is given by*

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

*Proof.* We only need to show that the square matrix $I - \gamma P^\pi$ is non-singular. Let $(\lambda_i)$ denote the eigenvalues of the matrix $P^\pi$. Since $P^\pi$ is a stochastic matrix (row sums are 1), then $|\lambda_i| \leq 1$. Now, the eignevalues of $I - \gamma P^\pi$ are $(1 - \gamma\lambda_i)$, and satisfy $|1 - \gamma\lambda_i| \geq 1 - \gamma > 0$. □

Combining Lemma 5.1 and Lemma 5.2, we obtain

**Proposition 5.3.** *Let* $\pi \in \Pi_{SD}$. *The value function* $\mathcal{V}^\pi = [\mathcal{V}^\pi(\mathtt{s})]$ *is the unique solution of equation* (5.2)*, given by*

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

Proposition 5.3 provides a closed-form formula for computing $\mathcal{V}^\pi$. However, for large systems, computing the inverse $(I - \gamma P^\pi)^{-1}$ may be computationally expensive. In that case, the following value iteration algorithm provides an alternative, iterative method for computing $\mathcal{V}^\pi$.

**Algorithm 5.1. *Fixed-policy value iteration***

1. *Let* $\mathcal{V}_0 = (\mathcal{V}_0(\mathtt{s}))_{\mathtt{s}\in\mathcal{S}}$ *be arbitrary.*

2. *For* $n = 0, 1, 2, \ldots$, *set*

$$\mathcal{V}_{n+1}(\mathtt{s}) = \mathtt{r}(\mathtt{s}, \pi(\mathtt{s})) + \gamma \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s}, \pi(\mathtt{s}))\mathcal{V}_n(\mathtt{s}'), \qquad \mathtt{s} \in \mathcal{S}$$

*or, equivalently,*

$$\mathcal{V}_{n+1} = r^\pi + \gamma P^\pi \mathcal{V}_n.$$

**Proposition 5.4 (Convergence of fixed-policy value iteration).** *We have* $\mathcal{V}_n \to \mathcal{V}^\pi$ *component-wise, that is,*

$$\lim_{n\to\infty} \mathcal{V}_n(\mathtt{s}) = \mathcal{V}^\pi(\mathtt{s}), \qquad \mathtt{s} \in \mathcal{S}.$$

*Proof.* Note first that

$$\begin{aligned}\mathcal{V}_1(\mathtt{s}) &= \mathtt{r}(\mathtt{s}, \pi(\mathtt{s})) + \gamma \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s}, \pi(\mathtt{s}))\mathcal{V}_0(\mathtt{s}') \\ &= \mathbb{E}^\pi(\mathtt{r}(\mathtt{s}_0, a_0) + \gamma\mathcal{V}_0(\mathtt{s}_1)|\mathtt{s}_0 = \mathtt{s}).\end{aligned}$$

Continuing similarly, we obtain that

$$\mathcal{V}_n(\mathtt{s}) = \mathbb{E}^\pi(\sum_{\mathtt{t}=0}^{n-1} \gamma^{\mathtt{t}} \mathtt{r}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) + \gamma^n \mathcal{V}_0(s_n)|\mathtt{s}_0 = \mathtt{s}).$$

Note that $\mathcal{V}_n(\mathtt{s})$ is the $n$-stage discounted return, with terminal reward $\mathtt{r}_n(\mathtt{s}_n) = \mathcal{V}_0(\mathtt{s}_n)$. Comparing with the definition of $\mathcal{V}^\pi$, we can see that

$$\mathcal{V}^\pi(\mathtt{s}) - \mathcal{V}_n(\mathtt{s}) = \mathbb{E}^\pi(\sum_{\mathtt{t}=n}^{\infty} \gamma^{\mathtt{t}} \mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) - \gamma^n \mathcal{V}_0(\mathtt{s}_n)|\mathtt{s}_0 = \mathtt{s}).$$

Denoting $\mathtt{R}_{\max} = \max_{\mathtt{s},\mathtt{a}} |\mathtt{r}(\mathtt{s},\mathtt{a})|$, $\bar{\mathcal{V}}_0 = \max_s |\mathcal{V}_0(\mathtt{s})|$ we obtain

$$|\mathcal{V}^\pi(\mathtt{s}) - \mathcal{V}_n(\mathtt{s})| \le \gamma^n (\frac{\mathtt{R}_{\max}}{1-\gamma} + \bar{\mathcal{V}}_0)$$

which converges to 0 since $\gamma < 1$. □

**Comments:**

- The proof provides an explicit bound on $|\mathcal{V}^\pi(\mathtt{s}) - \mathcal{V}_n(\mathtt{s})|$. It may be seen that the convergence is exponential, with rate $O(\gamma^n)$.
- Using vector notation, it may be seen that

$$\mathcal{V}_n = r^\pi + P^\pi r^\pi + \ldots + (P^\pi)^{n-1} r^\pi + (P^\pi)^n \mathcal{V}_0 = \sum_{\mathtt{t}=0}^{n-1} (P^\pi)^{\mathtt{t}} r^\pi + (P^\pi)^n \mathcal{V}_0.$$

  Similarly, $\mathcal{V}^\pi = \sum_{\mathtt{t}=0}^{\infty} (P^\pi)^{\mathtt{t}} r^\pi$.

**In summary:**

- Proposition 5.3 allows to compute $\mathcal{V}^\pi$ by solving a set of $|\mathcal{S}|$ linear equations.
- Proposition 5.4 computes $\mathcal{V}^\pi$ by an infinite recursion, that converges exponentially fast.

## 5.3 Overview: The Main DP Algorithms

We now return to the optimal planning problem defined in Section 5.1. Recall that $\mathcal{V}^*_\gamma(\mathtt{s}) = \sup_{\pi \in \Pi_{HS}} \mathcal{V}^\pi_\gamma(\mathtt{s})$ is the optimal discounted return. We further denote

$$\mathcal{V}^*(\mathtt{s}) \triangleq \mathcal{V}^*_\gamma(\mathtt{s}), \quad \mathtt{s} \in \mathcal{S},$$

and refer to $\mathcal{V}^*$ as the optimal value function. Depending on the context, we consider $\mathcal{V}^*$ either as a function $\mathcal{V}^* : \mathcal{S} \to \mathbb{R}$, or as a column vector $\mathcal{V}^* = [\mathcal{V}(\mathtt{s})]_{\mathtt{s}\in\mathcal{S}}$.

The following optimality equation provides an explicit characterization of the value function, and shows that an optimal stationary policy can easily be computed if the value function is known. (See the proof in Section 5.5.)

**Theorem 5.5** (**Bellman's Optimality Equation**). *The following statements hold:*

1. *$\mathcal{V}^*$ is the unique solution of the following set of (nonlinear) equations:*

$$\mathcal{V}(\mathtt{s}) = \max_{\mathtt{a}\in\mathcal{A}} \left\{ \mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\mathtt{a})\mathcal{V}(\mathtt{s}') \right\}, \quad \mathtt{s} \in \mathcal{S}. \tag{5.3}$$

2. *Any stationary policy $\pi^*$ that satisfies*

$$\pi^*(\mathtt{s}) \in \arg\max_{\mathtt{a}\in\mathcal{A}} \left\{ \mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\mathtt{a})\mathcal{V}(\mathtt{s}') \right\},$$

*is an optimal policy (for any initial state $\mathtt{s}_0 \in \mathcal{S}$).*

The optimality equation (5.3) is non-linear, and generally requires iterative algorithms for its solution. The main iterative algorithms are **value iteration** and **policy iteration**. In the following we provide the algorithms and the basic claims. Later in this chapter we formally prove the results regarding value iteration (Section 5.6) and policy iteration (Section 5.7).

**Algorithm 5.2.** ***Value Iteration (VI)***

1. *Let $\mathcal{V}_0 = (\mathcal{V}_0(\mathtt{s}))_{\mathtt{s}\in\mathcal{S}}$ be arbitrary.*
2. *For $n = 0, 1, 2, \ldots$, set*

$$\mathcal{V}_{n+1}(\mathtt{s}) = \max_{\mathtt{a}\in\mathcal{A}} \left\{ \mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\mathtt{a})\mathcal{V}_n(\mathtt{s}') \right\}, \quad \forall \mathtt{s} \in \mathcal{S}$$

**Theorem 5.6** (**Convergence of value iteration**). *We have $\lim_{n\to\infty}\mathcal{V}_n = \mathcal{V}^*$ (component-wise). The rate of convergence is exponential, at rate $O(\gamma^n)$.*

*Proof.* Using our previous results on value iteration for the finite-horizon problem, namely the proof of Proposition 5.4, it follows that

$$\mathcal{V}_n(\mathtt{s}) = \max_{\pi} \mathbb{E}^{\pi,\mathtt{s}}(\sum_{\mathtt{t}=0}^{n-1} \gamma^{\mathtt{t}} R_t + \gamma^n \mathcal{V}_0(\mathtt{s}_n)).$$

Comparing to the optimal value function

$$\mathcal{V}^*(\mathtt{s}) = \max_{\pi} \mathbb{E}^{\pi,\mathtt{s}}(\sum_{\mathtt{t}=0}^{\infty} \gamma^{\mathtt{t}} R_t),$$

it may be seen that that

$$|\mathcal{V}_n(\mathtt{s}) - \mathcal{V}^*(\mathtt{s})| \leq \gamma^n (\frac{\mathtt{R}_{\max}}{1-\gamma} + ||\mathcal{V}_0||_\infty).$$

As $\gamma < 1$, this implies that $\mathcal{V}_n$ converges to $\mathcal{V}^*_\gamma$ exponentially fast. □

The value iteration algorithm iterates over the value functions, with asymptotic convergence. The policy iteration algorithm iterates over stationary policies, with each new policy better than the previous one. This algorithm converges to the optimal policy in a finite number of steps.

**Algorithm 5.3.** ***Policy Iteration (PI)***

1. *Initialization: choose some stationary policy $\pi_0$.*
2. *For $k = 0, 1, \ldots$:*
   (a) *Policy evaluation: compute $\mathcal{V}^{\pi_k}$.*
   *(For example, use the explicit formula $\mathcal{V}^{\pi_k} = (I - \gamma P^{\pi_k})^{-1} r^{\pi_k}$.)*
   (b) *Policy Improvement: Compute $\pi_{k+1}$, a greedy policy with respect to $\mathcal{V}^{\pi_k}$:*
   $$\pi_{k+1}(\mathtt{s}) \in \arg\max_{\mathtt{a}\in\mathcal{A}} \left\{ \mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\mathtt{a}) \mathcal{V}^{\pi_k}(\mathtt{s}') \right\}, \quad \forall \mathtt{s} \in \mathcal{S}.$$
   (c) *Stop if $\pi_{k+1} = \pi_k$ (or if $\mathcal{V}^{\pi_k}$ satisfied the optimality equation), else continue.*

**Theorem 5.7** (**Convergence of policy iteration**). *The following statements hold:*

1. *Each policy $\pi_{k+1}$ is improving over the previous one $\pi_k$, in the sense that $\mathcal{V}^{\pi_{k+1}} \geq \mathcal{V}^{\pi_k}$ (component-wise).*
2. *$\mathcal{V}^{\pi_{k+1}} = \mathcal{V}^{\pi_k}$ if and only if $\pi_k$ is an optimal policy.*
3. *Consequently, since the number of stationary policies is finite, $\pi_k$ converges to the optimal policy after a finite number of steps.*

An additional solution method for DP planning relies on a Linear Programming formulation of the problem. We will provide additional details later in this Chapter, in Section **??**.

## 5.4 Contraction Operators

The basic proof methods of the DP results mentioned above rely on the concept of a *contraction operator*. We provide here the relevant mathematical background, and illustrate the contraction properties of some basic Dynamic Programming operators.

### 5.4.1 The contraction property

Recall that a norm $||\cdot||$ over $\mathbb{R}^n$ is a real-valued function $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$ such that, for any pair of vectors $x, y \in \mathbb{R}^d$ and scalar $a \in \mathbb{R}$,

1. $||ax|| = |a| \cdot ||x||$,
2. $||x+y|| \leq ||x|| + ||y||$,
3. $||x|| = 0$ only if $x = 0$.

Common examples are the p-norm $||x||_p = (\sum_{i=1}^d |x_i|^p)^{1/p}$ for $p \geq 1$, and in particular the Euclidean norm ($p = 2$). Here we will mostly use the max-norm:

$$||x||_\infty = \max_{1\leq i\leq d} |x_i|.$$

Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a vector-valued function over $\mathbb{R}^d$ ($d \geq 1$). We equip $\mathbb{R}^d$ with some norm $||\cdot||$, and refer to $T$ as an *operator* over $\mathbb{R}^d$. Thus, $T(v) \in \mathbb{R}^d$ for any $v \in \mathbb{R}^d$. We also denote $T^n(v) = T(T^{n-1}(v))$ for $n \geq 2$. For example, $T^2(v) = T(T(v))$.

**Definition 5.1.** *The operator $T$ is called a contraction operator if there exists $\beta \in (0,1)$ (the contraction coefficient) such that*

$$||T(v_1) - T(v_2)|| \leq \beta ||v_1 - v_2||,$$

*for all $v_1, v_2 \in \mathbb{R}^d$*

### 5.4.2 The Banach Fixed Point Theorem

The following celebrated result applies to contraction operators. While we quote the result for $\mathbb{R}^d$, we note that it applies in much greater generality to any Banach space (a complete normed space), or even to any complete metric space, with essentially the same proof.

**Theorem 5.8 (Banach's fixed point theorem).** *Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a contraction operator. Then*

1. *The equation $T(v) = v$ has a unique solution $\mathcal{V}^* \in \mathbb{R}^d$.*
2. *For any $v_0 \in \mathbb{R}^d$, $\lim_{n\to\infty} T^n(v_0) = \mathcal{V}^*$. In fact, $||T^n(v_0) - \mathcal{V}^*|| \le O(\beta^n)$, where $\beta$ is the contraction coefficient.*

*Proof.* Fix any $v_0$ and define $v_{n+1} = T(v_n)$. We will show that: (1) there exists a limit to the sequence, and (2) the limit is a fixed point of $T$.

**Existence of a limit $v^*$ of the sequence $v_n$**
We show that the sequence of $v_n$ is a cauchy sequence. We consider two elements $v_n$ and $v_{m+n}$ and bound the distance between them.

$$\begin{aligned}
\|v_{n+m} - v_n\| &= \|\sum_{k=0}^{m-1} v_{n+k+1} - v_{n+k}\| \\
&\le \sum_{k=0}^{m-1} \|v_{n+k+1} - v_{n+k}\| \quad (\textit{according to the triangle inequality}) \\
&= \sum_{k=0}^{m-1} \|T^{n+k}v_1 - T^{n+k}v_0\| \\
&\le \sum_{k=0}^{m-1} \beta^{n+k}\|v_1 - v_0\| \qquad (\textit{contraction } n+k \textit{ times}) \\
&= \frac{\beta^n(1-\beta^m)}{1-\beta}\|v_1 - v_0\|
\end{aligned}$$

Since the coefficient decreases as $n$ increases, for any $\epsilon > 0$ there exists $N > 0$ such that for all $n, m \ge N$, we have $\|\vec{v}_{n+m} - \vec{v}_n\| < \epsilon$. This implies that the sequence is a Cauchy sequence, and hence the sequence $v_n$ has a limit. Let us call this limit $v^*$. Next we show that $v^*$ is a fixed point of the operator $T$.

**$v^*$ is a fixed point**
We need to show that $T(v^*) = v^*$, or equivalently $\|T(v^*) - v^*\| = 0$.

$$\begin{aligned}
0 &\le \|T(v^*) - v^*\| \\
&\le \|T(v^*) - v_n\| + \|v_n - v^*\| \; (\textit{according to the triangle inequality}) \\
&= \|T(v^*) - T(v_{n-1})\| + \|v_n - v^*\| \\
&\le \beta\|\underbrace{v^* - v_{n-1}}_{\to 0}\| + \|\underbrace{v^n - v^*}_{\to 0}\|
\end{aligned}$$

Since $v^*$ is the limit of $v_n$, i.e., $\lim_{n\to\infty} \|\vec{v}_n - \vec{v^*}\| = 0$ hence

$$\|T\vec{v^*} - \vec{v^*}\| = 0.$$

Thus, $v^*$ is a fixed point of the operator $T$.

**Uniqueness of $\vec{v^*}$**
Assume that $T(v_1) = v_1$, and $T(v_2) = v_2$, and $v_1 \neq v_2$. Then

$$\|v_1 - v_2\| = \|T(v_1) - T(v_2)\| \leq \beta\|v_1 - v_2\|$$

Hence, this is in contradiction to $\beta < 1$. Therefore, $v^*$ is unique. □

### 5.4.3 The Dynamic Programming Operators

We next define the basic Dynamic Programming operators, and show that they are in fact contraction operators.

For a fixed stationary policy $\pi : \mathcal{S} \to \mathcal{A}$, define the fixed policy DP operator $T^\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as follows: For any $V = (V(\mathtt{s})) \in \mathbb{R}^{|\mathcal{S}|}$,

$$(T^\pi(V))(\mathtt{s}) = \mathtt{r}(\mathtt{s}, \pi(\mathtt{s})) + \gamma \sum\nolimits_{\mathtt{s}' \in \mathcal{S}} p(\mathtt{s}'|\mathtt{s}, \pi(\mathtt{s}))V(\mathtt{s}'), \quad \mathtt{s} \in \mathcal{S}$$

In our column-vector notation, this is equivalent to $T^\pi(V) = r^\pi + \gamma P^\pi V$.

Similarly, define the discounted-return **Dynamic Programming Operator** $T^*$ : $\mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as follows: For any $V = (V(\mathtt{s})) \in R^{|\mathcal{S}|}$,

$$(T^*(V))(\mathtt{s}) = \max_{\mathtt{a} \in \mathcal{A}} \left\{ \mathtt{r}(\mathtt{s}, \mathtt{a}) + \gamma \sum\nolimits_{\mathtt{s}' \in \mathcal{S}} p(\mathtt{s}'|\mathtt{s}, \mathtt{a})V(\mathtt{s}') \right\}, \quad \mathtt{s} \in \mathcal{S}$$

We note that $T^\pi$ is a linear operator, while $T^*$ is generally non-linear due to the maximum operation.

Let $||V||_\infty \overset{\Delta}{=} \max_{\mathtt{s}\in\mathcal{S}}|V(\mathtt{s})|$ denote the max-norm of $V$. Recall that $0 < \gamma < 1$.

**Theorem 5.9** (**Contraction property**). *The following statements hold:*

1. *$T^\pi$ is a $\gamma$-contraction operator with respect to the max-norm, namely $||T^\pi(\mathcal{V}_1) - T^\pi(\mathcal{V}_2)||_\infty \leq \gamma||\mathcal{V}_1 - \mathcal{V}_2||_\infty$ for all $\mathcal{V}_1, \mathcal{V}_2 \in \mathbb{R}^{|\mathcal{S}|}$.*

2. *Similarly, $T^*$ is an $\gamma$-contraction operator with respect to the max-norm.*

*Proof.* 1. Fix $\mathcal{V}_1, \mathcal{V}_2$. For every state $\mathtt{s}$,

$$\begin{aligned}
|(T^{\pi}(\mathcal{V}_1))(\mathtt{s}) - (T^{\pi}(\mathcal{V}_2))(\mathtt{s})| &= \left|\gamma \sum_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\pi(\mathtt{s}))[\mathcal{V}_1(\mathtt{s}') - \mathcal{V}_2(\mathtt{s}')]\right| \\
&\leq \gamma \sum_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\pi(\mathtt{s}))\, |\mathcal{V}_1(\mathtt{s}') - \mathcal{V}_2(\mathtt{s}')| \\
&\leq \gamma \sum_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\pi(\mathtt{s}))\, \|\mathcal{V}_1 - \mathcal{V}_2\|_\infty = \gamma \, \|\mathcal{V}_1 - \mathcal{V}_2\|_\infty \ .
\end{aligned}$$

Since this holds for every $\mathtt{s} \in \mathcal{S}$ the required inequality follows.

2. The proof here is more intricate due to the maximum operation. As before, we need to show that $|T^*(\mathcal{V}_1)(\mathtt{s}) - T^*(\mathcal{V}_2)(\mathtt{s})| \leq \gamma\|\mathcal{V}_1 - \mathcal{V}_2\|_\infty$. Fixing the state $\mathtt{s}$, we consider separately the positive and negative parts of the absolute value:

(a) Showing $T^*(\mathcal{V}_1)(\mathtt{s}) - T^*(\mathcal{V}_2)(\mathtt{s}) \leq \gamma\|\mathcal{V}_1 - \mathcal{V}_2\|_\infty$: Let $\bar{\mathtt{a}}$ denote an action that attains the maximum in $T^*(\mathcal{V}_1)(\mathtt{s})$, namely

$$\bar{\mathtt{a}} \in \arg\max_{\mathtt{a}\in\mathcal{A}} \left\{ \mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\mathtt{a})\mathcal{V}_1(\mathtt{s}') \right\}.$$

Then

$$\begin{aligned}
T^*(\mathcal{V}_1)(\mathtt{s}) &= \mathtt{r}(\mathtt{s},\bar{\mathtt{a}}) + \gamma \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\bar{\mathtt{a}})\mathcal{V}_1(\mathtt{s}') \\
T^*(\mathcal{V}_2)(\mathtt{s}) &\geq \mathtt{r}(\mathtt{s},\bar{\mathtt{a}}) + \gamma \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\bar{\mathtt{a}})\mathcal{V}_2(\mathtt{s}')
\end{aligned}$$

Since the same action $\bar{\mathtt{a}}$ appears in both expressions, we can now continue to show the inequality (a) similarly to 1. Namely,

$$\begin{aligned}
(T^{\pi}(\mathcal{V}_1))(\mathtt{s}) - (T^{\pi}(\mathcal{V}_2))(\mathtt{s}) &= \gamma \sum_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\pi(\mathtt{s}))\, (\mathcal{V}_1(\mathtt{s}') - \mathcal{V}_2(\mathtt{s}')) \\
&\leq \gamma \sum_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\pi(\mathtt{s}))\, \|\mathcal{V}_1 - \mathcal{V}_2\|_\infty = \gamma \, \|\mathcal{V}_1 - \mathcal{V}_2\|_\infty \ .
\end{aligned}$$

(b) Showing $T^*(\mathcal{V}_2)(\mathtt{s}) - T^*(\mathcal{V}_1)(\mathtt{s}) \leq \gamma\|\mathcal{V}_1 - \mathcal{V}_2\|_\infty$. By (a) we have

$$T^*(\mathcal{V}_2)(\mathtt{s}) - T^*(\mathcal{V}_1)(\mathtt{s}) \leq \gamma\|\mathcal{V}_2 - \mathcal{V}_1\|_\infty = \gamma\|\mathcal{V}_1 - \mathcal{V}_2\|_\infty.$$

The inequalities (a) and (b) together imply that $|T^*(\mathcal{V}_1)(\mathtt{s}) - T^*(\mathcal{V}_2)(\mathtt{s})| \leq \gamma\|\mathcal{V}_1 - \mathcal{V}_2\|_\infty$. Since this holds for any state $\mathtt{s}$, it follows that $||T^*(\mathcal{V}_1) - T^*(\mathcal{V}_2)||_\infty \leq \gamma\|\mathcal{V}_1 - \mathcal{V}_2\|_\infty$.

□

## 5.5 Proof of Bellman's Optimality Equation

We prove in this section Theorem 5.5, which is restated here:

**Theorem** (Bellman's Optimality Equation). *The following statements hold:*

1. *$\mathcal{V}^*$ is the unique solution of the following set of (nonlinear) equations:*

$$\mathcal{V}(\mathtt{s}) = \max_{\mathtt{a}\in\mathcal{A}} \left\{ \mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\mathtt{a})\mathcal{V}(\mathtt{s}') \right\}, \quad \mathtt{s}\in\mathcal{S}. \tag{5.4}$$

2. *Any stationary policy $\pi^*$ that satisfies*

$$\pi^*(\mathtt{s}) \in \arg\max_{\mathtt{a}\in\mathcal{A}} \left\{ \mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\mathtt{a})\mathcal{V}(\mathtt{s}') \right\},$$

   *is an optimal policy (for any initial state $\mathtt{s}_0\in\mathcal{S}$).*

We observe that the Optimality equation in part 1 is equivalent to $V = T^*(V)$ where $T^*$ is the optimal DP operator from the previous section, which was shown to be a contraction operator with coefficient $\gamma$. The proof also uses the value iteration property of Theorem 5.6.

***Proof of Theorem 5.5:*** We prove each part.

1. As $T^*$ is a contraction operator, existence and uniqueness of the solution to $V = T^*(V)$ follows from the Banach fixed point theorem (Theorem 5.8). Let $\hat{V}$ denote that solution. It also follows by that theorem (Theorem 5.8) that $(T^*)^n(\mathcal{V}_0) \to \hat{V}$ for any $\mathcal{V}_0$. By Theorem 5.6 we have that $(T^*)^n(\mathcal{V}_0) \to \mathcal{V}^*$, hence $\hat{V} = \mathcal{V}^*$, so that $\mathcal{V}^*$ is indeed the unique solution of $V = T^*(V)$.

2. By definition of $\pi^*$ we have

$$T^{\pi^*}(\mathcal{V}^*) = T^*(\mathcal{V}^*) = \mathcal{V}^*,$$

   where the last equality follows from part 1. Thus the optimal value function satisfied the equation $T^{\pi^*}(\mathcal{V}^*) = \mathcal{V}^*$. But we already know (from Prop. 5.4) that $\mathcal{V}^{\pi^*}$ is the unique solution of that equation, hence $\mathcal{V}^{\pi^*} = \mathcal{V}^*$. This implies that $\pi^*$ achieves the optimal value (for any initial state), and is therefore an optimal policy as stated.

□

## 5.6 Value Iteration (VI)

The value iteration algorithm allows to compute the optimal value function $\mathcal{V}^*$ iteratively to any required accuracy. The Value Iteration algorithm (Algorithm 5.2) can be stated as follows:

1. Start with any initial value function $\mathcal{V}_0 = (\mathcal{V}_0(\mathtt{s}))$.

2. Compute recursively, for $n = 0, 1, 2, \ldots$,

$$\mathcal{V}_{n+1}(\mathtt{s}) = \max_{\mathtt{a}\in\mathcal{A}} \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\mathtt{a})[\mathtt{r}(\mathtt{s},\mathtt{a},\mathtt{s}') + \gamma\mathcal{V}_n(\mathtt{s}')], \quad \forall \mathtt{s}\in\mathcal{S}.$$

3. Apply a stopping rule to obtain a required accuracy (see below).

In terms of the DP operator $T^*$, value iteration is simply stated as:

$$\mathcal{V}_{n+1} = T^*(\mathcal{V}_n), \quad n \geq 0.$$

Note that the number of operations for each iteration is $O(|\mathcal{A}| \cdot |\mathcal{S}|^2)$. Theorem 5.6 states that $\mathcal{V}_n \to \mathcal{V}^*$, exponentially fast.

### 5.6.1 Error bounds and stopping rules:

While we showed an exponential convergence rate, it is important to have a criteria that would depend only on the observed quantities.

**Lemma 5.10.** *If $\|\mathcal{V}_{n+1}-\mathcal{V}_n\|_\infty < \epsilon\cdot\frac{1-\gamma}{2\gamma}$ then $\|\mathcal{V}_{n+1}-\mathcal{V}^*\|_\infty < \frac{\varepsilon}{2}$ and $|\mathcal{V}^{\pi_{n+1}}-\mathcal{V}^*| \leq \varepsilon$, where $\pi_{n+1}$ is the greedy policy w.r.t. $\mathcal{V}_{n+1}$.*

*Proof.* Assume that $\|\mathcal{V}_{n+1}-\mathcal{V}_n\| < \varepsilon\cdot\frac{1-\gamma}{2\gamma}$, and we show that $\|\mathcal{V}^{\pi_{n+1}}-\mathcal{V}^*\| < \varepsilon$, which would make the policy $\pi_{n+1}$ $\varepsilon$-optimal. We bound the difference between $\mathcal{V}^{\pi_{n+1}}$ and $\mathcal{V}^*$. (All the norms are max-norm.) We consider the following:

$$\|\mathcal{V}^{\pi_{n+1}} - \mathcal{V}^*\| \leq \|\mathcal{V}^{\pi_{n+1}} - \mathcal{V}_{n+1}\| + \|\mathcal{V}_{n+1} - \mathcal{V}^*\| \tag{5.5}$$

We now bound each part of the sum separately:

$$\begin{aligned}\|\mathcal{V}^{\pi_{n+1}} - \mathcal{V}_{n+1}\| &= \|T^{\pi_{n+1}}(\mathcal{V}^{\pi_{n+1}}) - \mathcal{V}_{n+1}\| \qquad (\textit{because } \mathcal{V}^{\pi_{n+1}} \textit{ is the fixed point of } T^{\pi_{n+1}})\\ &\leq \|T^{\pi_{n+1}}(\mathcal{V}^{\pi_{n+1}}) - T^*(\mathcal{V}_{n+1})\| + \|T^*(\mathcal{V}_{n+1}) - \mathcal{V}_{n+1}\|\end{aligned}$$

Since $\pi_{n+1}$ is maximal over the actions using $\mathcal{V}_{n+1}$, it implies that $T^{\pi_{n+1}}(\mathcal{V}_{n+1}) = T^*(\mathcal{V}_{n+1})$ and we conclude that:

$$\begin{aligned}\|\mathcal{V}^{\pi_{n+1}} - \mathcal{V}_{n+1}\| &\leq \|T^{\pi_{n+1}}(\mathcal{V}^{\pi_{n+1}}) - T^{\pi_{n+1}}(\mathcal{V}_{n+1})\| + \|T^*(\mathcal{V}_{n+1}) - T^*(\mathcal{V}_n)\| \\ &\leq \gamma\|\mathcal{V}^{\pi_{n+1}} - \mathcal{V}_{n+1}\| + \gamma\|\mathcal{V}_{n+1} - \mathcal{V}_n\|\end{aligned}$$

Rearranging, this implies that,

$$\|\mathcal{V}^{\pi_{n+1}} - \mathcal{V}_{n+1}\| \leq \frac{\gamma}{1-\gamma}\|\mathcal{V}_{n+1} - \mathcal{V}_n\| < \frac{\gamma}{1-\gamma} \cdot \epsilon \cdot \frac{1-\gamma}{2\gamma} = \frac{\epsilon}{2}$$

For the second part of the sum we derive similarly that:

$$\|\mathcal{V}_{n+1} - \mathcal{V}^*\| \leq \frac{\gamma}{1-\gamma}\|\mathcal{V}_{n+1} - \mathcal{V}_n\| < \frac{\gamma}{1-\gamma} \cdot \epsilon \cdot \frac{1-\gamma}{2\gamma} = \frac{\epsilon}{2}$$

Returning to inequality (5.5), it follows:

$$\|\mathcal{V}^{\pi_{n+1}} - \mathcal{V}^*\| \leq \frac{2\gamma}{1-\gamma}\|\mathcal{V}_{n+1} - \mathcal{V}_n\| < \epsilon$$

Therefore the selected policy $\pi_{n+1}$ is $\epsilon$-optimal. $\square$

## 5.7 Policy Iteration (PI)

The policy iteration algorithm, introduced by Howard (1960), computes an optimal policy $\pi^*$ in a finite number of steps. This number is typically small (on the same order as $|\mathcal{S}|$).

The basic principle behind Policy Iteration is Policy Improvement. Let $\pi$ be a stationary policy, and let $\mathcal{V}^\pi$ denote its value function. A stationary policy $\bar{\pi}$ is called $\pi$- improving if it is a greedy policy with respect to $\mathcal{V}^\pi$, namely

$$\bar{\pi}(\mathtt{s}) \in \arg\max_{\mathtt{a}\in\mathcal{A}} \left\{\mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \sum\nolimits_{\mathtt{s}'\in\mathcal{S}} p(\mathtt{s}'|\mathtt{s},\mathtt{a})\mathcal{V}^\pi(\mathtt{s}')\right\}, \quad \mathtt{s} \in \mathcal{S}.$$

**Lemma 5.11** (Policy Improvement). *Let $\pi$ be a stationary policy and $\bar{\pi}$ be a $\pi$-improving policy. We have $\mathcal{V}^{\bar{\pi}} \geq \mathcal{V}^\pi$ (component-wise), and $\mathcal{V}^{\bar{\pi}} = \mathcal{V}^\pi$ if and only if $\pi$ is an optimal policy.*

*Proof.* Observe first that

$$\mathcal{V}^\pi = T^\pi(\mathcal{V}^\pi) \leq T^*(\mathcal{V}^\pi) = T^{\bar{\pi}}(\mathcal{V}^\pi)$$

The first equality follows since $\mathcal{V}^\pi$ is the value function for the policy $\pi$, the inequality follows because of the maximization in the definition of $T^*$, and the last equality by definition of the improving policy $\bar{\pi}$.

It is easily seen that $T^\pi$ is a monotone operator (for any policy $\pi$), namely $\mathcal{V}_1 \leq \mathcal{V}_2$ implies $T^\pi(\mathcal{V}_1) \leq T^\pi(\mathcal{V}_2)$. Applying $T^{\bar{\pi}}$ repeatedly to both sides of the above inequality $\mathcal{V}^\pi \leq T^{\bar{\pi}}(\mathcal{V}^\pi)$ therefore gives

$$\mathcal{V}^\pi \leq T^{\bar{\pi}}(\mathcal{V}^\pi) \leq (T^{\bar{\pi}})^2(\mathcal{V}^\pi) \leq \cdots \leq \lim_{n\to\infty} (T^{\bar{\pi}})^n(\mathcal{V}^\pi) = \mathcal{V}^{\bar{\pi}},$$

where the last equality follows by Theorem 5.6 This establishes the first claim.

We now show that $\pi$ is optimal if and only if $\mathcal{V}^{\bar{\pi}} = \mathcal{V}^\pi$. We showed that $\mathcal{V}^{\bar{\pi}} \geq \mathcal{V}^\pi$. If $\mathcal{V}^{\bar{\pi}} > \mathcal{V}^\pi$ then clearly $\pi$ is not optimal. Assume that $\mathcal{V}^{\bar{\pi}} = \mathcal{V}^\pi$. We have the following identities:

$$\mathcal{V}^\pi = \mathcal{V}^{\bar{\pi}} = T^{\bar{\pi}}(\mathcal{V}^{\bar{\pi}}) = T^{\bar{\pi}}(\mathcal{V}^\pi) = T^*(\mathcal{V}^{\bar{\pi}}) = T^*(\mathcal{V}^\pi)$$

where the first equality is by our assumption. The second equality follows since $\mathcal{V}^{\bar{\pi}}$ is the fixed point of its operator $T^{\bar{\pi}}$. The third follows since we assume that $\mathcal{V}^{\bar{\pi}} = \mathcal{V}^\pi$. The last equality follows since $T^{\bar{\pi}}$ and $T^*$ are identical on $\mathcal{V}^\pi$.

We have established that: $\mathcal{V}^\pi = T^*(\mathcal{V}^\pi)$, and hence $\mathcal{V}^\pi$ and $\pi$ is a fixed point of $T^*$ and therefore, by Theorem 5.5, policy $\pi$ is optimal. □

The policy iteration algorithm performs successive rounds of policy improvement, where each policy $\pi_{k+1}$ improves the previous one $\pi_k$. Since the number of stationary policies is bounded, so is the number of strict improvements, and the algorithm must terminate with an optimal policy after a finite number of iterations.

In terms of computational complexity, Policy Iteration requires $O(|\mathcal{A}| \cdot |\mathcal{S}|^2 + |\mathcal{S}|^3)$ operations per iteration, while Value Iteration requires $O(|\mathcal{A}| \cdot |\mathcal{S}|^2)$ per iteration. However, in many cases the Policy Iteration has a smaller number of iteration than Value Iteration, as we show in the next section. Another consideration is that the number of iterations of Value Iteration increases as the discount factor $\gamma$ approaches 1, while the number of policies (which upper bound the number of iterations of Policy Iteration) is independent from $\gamma$.

## 5.8 A Comparison between VI and PI Algorithms

In this section we will compare the convergence rate of the VI and PI algorithms. We show that, assuming that the two algorithms begin with the same approximated value, the PI algorithm converges in less iterations.

**Theorem 5.12.** *Let $\{VI_n\}$ be the sequence of values created by the VI algorithm (where $VI_{n+1} = T^*(VI_n)$) and let $\{PI_n\}$ be the sequence of values created by PI algorithm, i.e., $PI_n = \mathcal{V}^{\pi_n}$. If $VI_0 = PI_0$, then for all $n$ we have $VI_n \leq PI_n \leq \mathcal{V}^*$.*

*Proof.* The proof is by induction on $n$.
*Induction Basis:* By construction $VI_0 = PI_0$. Since $PI_0 = \mathcal{V}^{\pi_0}$, it is clearly bounded by $\mathcal{V}^*$.
*Induction Step:* Assume that $VI_n \leq PI_n$. For $VI_{n+1}$ we have,

$$VI_{n+1} = T^*(VI_n) = T^{\pi'}(VI_n),$$

where $\pi'$ is the greedy policy w.r.t. $VI_n$, i.e.,

$$\pi'(\mathtt{s}) \in \arg\max_{a \in A}\{\mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \sum_{\mathtt{s}' \in \mathcal{S}} p(\mathtt{s}'|\mathtt{s},\mathtt{a}) VI_n(\mathtt{s}')\}.$$

Since $VI_n \leq PI_n$, and $T^{\pi'}$ is monotonic it follows that:

$$T^{\pi'}(VI_n) \quad \leq \quad T^{\pi'}(PI_n)$$

Since $T^*$ is upper bounding any $T^\pi$:

$$T^{\pi'}(PI_n) \quad \leq \quad T^*(PI_n)$$

The policy determined by PI algorithm in iteration $n+1$ is $\pi_{n+1}$ and we have:

$$T^*(PI_n) = T^{\pi_{n+1}}(PI_n)$$

From the definition of $\pi_{n+1}$, we have

$$T^{\pi_{n+1}}(PI_n) \quad \leq \quad \mathcal{V}^{\pi_{n+1}} = PI_{n+1}$$

Therefore, $VI_{n+1} \leq PI_{n+1}$. Since $PI_{n+1} = \mathcal{V}^{\pi_{n+1}}$, it implies that $PI_{n+1} \leq \mathcal{V}^*$. □

## 5.9 Some Variants on Value Iteration and Policy Iteration

### 5.9.1 Value Iteration - Gauss Seidel Iteration

In the standard value iteration: $\mathcal{V}_{n+1} = T^*(\mathcal{V}_n)$, the vector $\mathcal{V}_n$ is held fixed while all entries of $\mathcal{V}_{n+1}$ are updated. An alternative is to update each element $\mathcal{V}_n(\mathtt{s})$ of that vector as to $\mathcal{V}_{n+1}(\mathtt{s})$ as soon as the latter is computed, and continue the calculation with the new value. This procedure is guaranteed to be "as good" as the standard one, in some sense, and often speeds up convergence.

### 5.9.2 Asynchronous Value Iteration

Here, in each iteration $\mathcal{V}_n \Rightarrow \mathcal{V}_{n+1}$, only a subset of the entries of $\mathcal{V}_n$ (namely, a subset of all states) is updated. It can be shown that if each state is updated infinitely often, then $\mathcal{V}_n \rightarrow \mathcal{V}^*$. Asynchronous update can be used to focus the computational effort on "important" parts of a large-state space.

### 5.9.3 Modified (a.k.a. Generalized or Optimistic) Policy Iteration

This scheme combines policy improvement steps with value iteration for policy evaluation. This way the requirement for exact policy evaluation (computing $\mathcal{V}^{\pi_k} = (I - \gamma P^{\pi_k})^{-1} r^{\pi_k}$) is avoided.

The procedure starts with some initial value vector $\mathcal{V}_0$, and iterates as follows:

- Greedy policy computation:

  Compute $\pi_k \in \arg\max_\pi T^\pi(\mathcal{V}_k)$, a greedy policy with respect to $\mathcal{V}_k$.

- Partial value iteration:

  Perform $m_k$ steps of value iteration, $\mathcal{V}_{k+1} = (T_\gamma^{\pi_k})^{m_k}(\mathcal{V}_k)$.

This algorithm guarantees convergence of $\mathcal{V}_k$ to $\mathcal{V}^*$.

## 5.10 Linear Program

In this section we will use linear programming to derive the optimal policy for discounted return. We will extend the linear programming given in Section 4.5 from Finite Horizon return to discounted return, but the derivation is similar in spirit. We will see that both the primal and dual program will play an important part in defining the optimal policy. We will fix an initial state $\mathtt{s}_0$ and compute the optimal policy for it.

We will start with the primal linear program, which will compute the optimal policy. For each state $\mathtt{s}$ and action $\mathtt{a}$ we will have a variable $x(\mathtt{s}, \mathtt{a})$ that will indicate the discounted fraction of time we are at state $\mathtt{s}$ and perform action $\mathtt{a}$.

To better understand what we mean by the "discounted fraction of time" consider a fixed policy $\pi$ and a trajectory $(\mathtt{s}_0, \ldots)$ generated by $\pi$. Define $X^\pi(\mathtt{s}, \mathtt{a}) = \sum_\mathtt{t} \mathbb{I}(\mathtt{s}_\mathtt{t} = \mathtt{s}, \mathtt{a}_\mathtt{t} = \mathtt{a})$, which is a random variable. We are interested in $x^\pi(\mathtt{s}, \mathtt{a}) = \mathbb{E}[X^\pi(\mathtt{s}, \mathtt{a})]$ which is the expected discounted fraction of time policy $\pi$ is in state $\mathtt{s}$ and performs action $\mathtt{a}$. The goal of our linear program is to compute $x^{\pi^*}$ for the optimal policy $\pi^*$.

Our main constraint will be a flow constraint, stating that the discounted fraction of time we reach state $\mathtt{s}$ upper bounds the discounted fraction of time we exit it, times the discounted factor. Formally, for $\mathtt{s} \in \mathcal{S}$,

$$\mathbb{I}(\mathtt{s} = \mathtt{s}_0) + \sum_{\mathtt{a}} x(\mathtt{s}, \mathtt{a}) \leq \gamma \sum_{\mathtt{s}', \mathtt{a}'} x(\mathtt{s}', \mathtt{a}') p(\mathtt{s}|\mathtt{s}'\mathtt{a}').$$

Note that if we sum the inequalities over all states, we have

$$1 + \sum_{\mathtt{s},\mathtt{a}} x(\mathtt{s}, \mathtt{a}) \leq \gamma \sum_{\mathtt{s}',\mathtt{a}'} x(\mathtt{s}', \mathtt{a}') \sum_{\mathtt{s}} p(\mathtt{s}|\mathtt{s}'\mathtt{a}') = \gamma \sum_{\mathtt{s}',\mathtt{a}'} x(\mathtt{s}', \mathtt{a}').$$

which implies that $\sum_{\mathtt{s},\mathtt{a}} x(\mathtt{s}, \mathtt{a}) \leq 1/(1-\gamma)$, as we should expect. Namely, in each time we are in some state, therefore the sum over states should be $\sum_{\mathtt{t}} \gamma^{\mathtt{t}} = 1/(1-\gamma)$.

The discounted return, which we would like to maximize, is $\mathbb{E}[\sum_{\mathtt{t}} \gamma^{\mathtt{t}} \mathtt{r}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}})]$. We can regroup the sum by state and action and have $\sum_{\mathtt{s},\mathtt{a}} \mathbb{E}[\sum_{\mathtt{t}} \gamma^{\mathtt{t}} \mathtt{r}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) \mathbb{I}(\mathtt{s}_{\mathtt{t}} = \mathtt{s}, \mathtt{a}_{\mathtt{t}} = \mathtt{a})]$, which is equivalent to $\sum_{\mathtt{s},\mathtt{a}} \mathtt{r}(\mathtt{s}, \mathtt{a}) \mathbb{E}[\sum_{\mathtt{t}} \gamma^{\mathtt{t}} \mathbb{I}(\mathtt{s}_{\mathtt{t}} = \mathtt{s}, \mathtt{a}_{\mathtt{t}} = \mathtt{a})]$. Now our variable are $x(\mathtt{s}, \mathtt{a}) = \mathbb{E}[\sum_{\mathtt{t}} \gamma^{\mathtt{t}} \mathbb{I}(\mathtt{s}_{\mathtt{t}} = \mathtt{s}, \mathtt{a}_{\mathtt{t}} = \mathtt{a})]$, and the expected return would be

$$\sum_{\mathtt{s},\mathtt{a}} \mathtt{r}(\mathtt{s}, \mathtt{a}) x(\mathtt{s}, \mathtt{a})$$

The resulting linear program is the following.

$$\begin{aligned}
\max_{x(\mathtt{s},\mathtt{a})} \quad & \sum_{\mathtt{s},\mathtt{a}} \mathtt{r}(\mathtt{s}, \mathtt{a}) x(\mathtt{s}, \mathtt{a}) \\
& \text{such that} \\
& \mathbb{I}(\mathtt{s} = \mathtt{s}_0) + \sum_{\mathtt{a}} x(\mathtt{s}, \mathtt{a}) \leq \gamma \sum_{\mathtt{s}',\mathtt{a}'} x(\mathtt{s}', \mathtt{a}') p(\mathtt{s}|\mathtt{s}'\mathtt{a}') && \forall \mathtt{s} \in \mathcal{S}, \mathtt{a} \in \mathcal{A}, \\
& x(\mathtt{s}, \mathtt{a}) \geq 0 && \forall \mathtt{s} \in \mathcal{S}, \mathtt{a} \in \mathcal{A}, \\
& \sum_{\mathtt{a}} x_0(\mathtt{s}_0, \mathtt{a}) = 1 && \forall \mathtt{a} \in \mathcal{A}, \\
& x_0(\mathtt{s}, \mathtt{a}) = 0, && \forall \mathtt{s} \in \mathcal{S}, \mathtt{s} \neq \mathtt{s}_0
\end{aligned}$$

Given the primal linear program we can derive the dual linear program.

$$\min_{z(\mathtt{s})} \; z_0(\mathtt{s}_0)$$

such that

$$z(\mathtt{s}) \geq \mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \sum_{\mathtt{s}'} z(\mathtt{s}')p(\mathtt{s}'|\mathtt{s},\mathtt{a}), \qquad \forall \mathtt{s} \in \mathcal{S}, \mathtt{a} \in \mathcal{A},$$

.

One can identify the dual random variables $z(\mathtt{s})$ with the optimal vale function $\mathcal{V}(\mathtt{s})$. At the optimal solution of the dual linear program one can show that we have

$$z(\mathtt{s}) = \max_{\mathtt{a}} \big\{\mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \sum_{\mathtt{s}'} z(\mathtt{s}')p_{\mathtt{t}}(\mathtt{s}'|\mathtt{s},\mathtt{a})\big\}, \qquad \forall \mathtt{s} \in \mathcal{S},$$

which are the familiar Bellman optimality equations.

# Chapter 6

# Stochastic shortest paths

This class of problems provides a natural extension of the standard shortest-path problem to stochastic settings. When we view Stochastic Shortest Paths (SSP) as an extension of the graph theoretic notion of shortest paths, we can motivate it by having the edges no completely deterministic, but rather having a probability of ending a different state. Probably a better view, is to think of the edges as general actions, which induce a distribution over the next state. The goal state can be either a single state or a set of states, both notions would be equivalent.

The SSP problem includes an important sub-category, which is *episodic MDP*. In an episodic MDP we are guarantee to complete the episode in (expected) finite time, regardless of the policy we employ. This will not be true of a general SSP, as some policies

Some conditions on the system dynamics and reward function must be imposed for the problem to be well posed (e.g., that a goal state may be reached with probability one). Such problems are known as stochastic shortest path problems, or also episodic planning problems.

## 6.1 Definition

Stochastic Shortest Path is an important class of planning problems, where the time horizon is not set beforehand, but rather the problem continues until a certain event occurs. This event can be defined as reaching some goal state. Let $\mathcal{S}_G \subset \mathcal{S}$ define the set of *goal states*.

Define

$$\tau = \inf\{t \geq 0 : \mathtt{s_t} \in \mathcal{S}_G\}$$

as the first time in which a goal state is reached.

The total expected return for Stochastic Shortest Path problem is defined as:

$$\mathcal{V}^{\pi}_{ssp}(\mathtt{s}) = \mathbb{E}^{\pi,\mathtt{s}}(\sum_{\mathtt{t}=0}^{\tau-1} \mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) + \mathtt{r}_G(\mathtt{s}_\tau))$$

Here $\mathtt{r}_G(\mathtt{s}),\ \mathtt{s} \in \mathcal{S}_G$ specified the reward at goal states. Note that the length of the run $\tau$ is a random variable.

### 6.1.1 Cost versus reward

While we can always switch the optimization from minimizing costs to maximizing rewards, in this case the two would have a conceptually different optimal policies. If we have non-negative costs, then the optimal policy would try to reach the goal states as fast as possible. (Actually, would minimize the expected cost in reaching the goal states.) When we have non-negative rewards, the optimal policy would try to avoid the goal states.

When we are considering a Deterministic Decision Process we can illustrate the difference. Recall that DMP is simply a graph where the actions are traversing edges. For costs, we are looking for the minimum cost path to the goal. For rewards, we are looking for a cycle with positive costs. (In Chapter 2, we derive an algorithm for the average reward for DDP.) Note that if there is a cycle of negative cost, then the shortest path is not well define (we can get infinite negative cost). Similarly, if we have a positive reward cycle we can get infinite reward.

## 6.2 Relationship to other models

### 6.2.1 Finite Horizon Return

Stochastic shortest path includes, naturally, the finite horizon case. This can be shown by creating a leveled MDP where at each time step we move to the next level and terminate at level $\mathtt{T}$. Specifically, we define a new state space $\mathcal{S}' = \mathcal{S} \times \mathbb{T}$, transition function $p((\mathtt{s}', i+1)|\mathtt{s}, \mathtt{a}) = p(\mathtt{s}'|\mathtt{s}, \mathtt{a})$, and goal states $\mathcal{S}_G = \{(\mathtt{s}, \mathtt{T}) : \mathtt{s} \in \mathcal{S}\}$.

### 6.2.2 Discounted infinite return

Stochastic shortest path includes also the discounted infinite horizon. To see that, add a new goal state, and from each state with probability $1-\gamma$ jump to the goal state and terminate. The expected return of a policy would be the same in both models. Specifically, we add a state $\mathtt{s}_G$, such that $p(\mathtt{s}_G|\mathtt{s},\mathtt{a}) = 1-\gamma$, for any state $\mathtt{s} \in \mathcal{S}$ and action $\mathtt{a} \in \mathcal{A}$ and $p(\mathtt{s}'|\mathtt{s},\mathtt{a}) = \gamma p(\mathtt{s}'|\mathtt{s},\mathtt{a})$. The probability that we do not terminate by time $\mathtt{t}$ is exactly $\gamma^{\mathtt{t}}$. Therefore the expected return is $\sum_{t=1}^{\infty} \gamma^t \mathtt{r}(\mathtt{s}_t, \mathtt{a}_t)$ which is identical to the discounted return.

## 6.3 Proper versus improper policies

Episodic MDP

### 6.3.1 Proper policies

Show that if we add $\epsilon > 0$ to all the cost, for sufficiently small $\epsilon$ we get the optimal proper policy.

## 6.4 Belman Operator

Add assumption that any improper policy has infinite cost.

Show that the Belman operator converge.

# Chapter 7

# Average cost

# Chapter 8

# Tabular learning: Model based

Until now we looked at planning problems, where we are given a complete model of the MDP, and the goal is to either evaluate a given policy or compute the optimal policy. In this chapter we will start looking at learning problems, where we need to learn from interaction. This chapter will concentrate on *model based* learning, where the main goal is to learn an accurate model of the MDP and use it. In following chapters we will look at *model free* learning, where we learn a value function or a policy without recovering the actual underlying model.

## 8.1 Effective horizon of discounted return

Before we start looking at the learning setting, we will show a "reduction" from discounted return to finite horizon return. The main issue will be to show that the discounted return has an *effective horizon* such that rewards beyond it have a negligible effect on the discounted return.

**Theorem 8.1.** *Given a discount factor $\gamma$, the discounted return in the first $\mathtt{T} = \frac{1}{1-\gamma}\log\frac{\mathtt{R}_{\max}}{\varepsilon(1-\gamma)}$ time steps, is within $\varepsilon$ of the total discounted return.*

*Proof.* Recall that the rewards are $\mathtt{r_t} \in [0, \mathtt{R}_{\max}]$. Fix an infinite sequence of rewards $(\mathtt{r}_0, \dots, \mathtt{r_t}, \dots)$. We would like to consider the following difference:

$$\sum_{\mathtt{t}=1}^{\infty} \mathtt{r_t}\gamma^{\mathtt{t}} - \sum_{\mathtt{t}=0}^{\mathtt{T}-1} \mathtt{r_t}\gamma^{\mathtt{t}} = \sum_{\mathtt{t}=\mathtt{T}}^{\infty} \mathtt{r_t}\gamma^{\mathtt{t}} \leq \frac{\gamma^{\mathtt{T}}}{1-\gamma}\mathtt{R}_{\max},$$

We want this difference to be bounded by $\varepsilon$, hence

$$\frac{\gamma^{\mathtt{T}}}{1-\gamma}\mathtt{R}_{\max} \leq \varepsilon\ .$$

This is equivalent to,

$$\mathtt{T}\log(1/\gamma) \geq \log \frac{\mathtt{R}_{\max}}{\varepsilon(1-\gamma)} .$$

Since $\log(1+x) \leq x$, we can bound $\log(1/\gamma) = \log(1 + \frac{1-\gamma}{\gamma}) \leq \frac{1-\gamma}{\gamma}$. Since $\gamma < 1$, we have that $\frac{\gamma}{1-\gamma} \leq \frac{1}{1-\gamma}$ and hence it is sufficient to have $\mathtt{T} \geq \frac{1}{1-\gamma} \log \frac{\mathtt{R}_{\max}}{\varepsilon(1-\gamma)}$, and the theorem follows. □

## 8.2 Off-Policy Model-Based Learning

In the off-policy setting we would have access to previous executed trajectories, and we would like to use them to learn. Naturally, we will have to make some assumption about the trajectories. Intuitively, we will need to assume that they are sufficiently exploratory.

We will decompose the trajectories to quadruples, which are composed of $(\mathtt{s}, \mathtt{a}, \mathtt{r}, \mathtt{s}')$ where $\mathtt{r}$ is sampled from $\mathtt{R}(\mathtt{s}, \mathtt{a})$ and $\mathtt{s}'$ is sampled from $p(\cdot|\mathtt{s}, \mathtt{a})$.

Our goal is to output an MDP $(\mathcal{S}, \mathcal{A}, \widehat{\mathtt{r}}, \widehat{p})$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\widehat{\mathtt{r}}(\mathtt{s}, \mathtt{a})$ is the approximate expected reward of $\mathtt{R}(\mathtt{s}, \mathtt{a}) \in [0, \mathtt{R}_{\max}]$, and $\widehat{p}(\mathtt{s}'|\mathtt{s}, \mathtt{a})$ is the approximate probability of reaching state $\mathtt{s}'$ when we are in state $\mathtt{s}$ and doing action $\mathtt{a}$. Intuitively, we would like to have the approximate model and the true model a similar expected value for any policy.

### 8.2.1 Mean estimation

We start with a basic mean estimation problem, which appears in many settings including supervised learning. Suppose we are given access to a random variable $\mathtt{R} \in [0, 1]$ and would like to approximate its mean $\mu = \mathbb{E}[\mathtt{R}]$. We observe $m$ samples of $\mathtt{R}$, which are $R_1, \ldots, R_m$, and compute their observed mean $\widehat{\mu} = \frac{1}{m}\sum_{i=1}^{m} R_i$.

By the law of large numbers we know that when $m$ goes to infinity we have that $\widehat{\mu}$ converges to $\mu$. We would like to have concrete finite convergence bounds, mainly to derive the value of $m$ as a function of the desired accuracy $\varepsilon$. For this we use concentration bounds (known as Chernoff-Hoffding bounds). The bounds have both an additive form and a multiplicative form, given as follows:

**Lemma 8.2** (Chernoff-Hoffding)**.** *Let $R_1, \ldots, R_m$ be $m$ i..i.d. samples of a random variable $\mathtt{R} \in [0, 1]$. Let $\mu = \mathbb{E}[\mathtt{R}]$ and $\widehat{\mu} = \frac{1}{m}\sum_{i=1}^{m} R_i$. For any $\varepsilon \in (0, 1)$ we have,*

$$\Pr[|\mu - \widehat{\mu}| \geq \varepsilon] \leq 2e^{-2\varepsilon^2 m}$$

*In addition,*

$$\Pr[\widehat{\mu} \le (1-\varepsilon)\mu] \le e^{-\varepsilon^2 m/2} \qquad \textit{and} \qquad \Pr[\widehat{\mu} \ge (1+\varepsilon)\mu] \le e^{-\varepsilon^2 m/3}$$

We will refer to the first bound as *additive* and the second set of bound as *multiplicative.*

Using the additive bound of Lemma 8.2, we have

**Corollary 8.3.** *Let $R_1, \ldots, R_m$ be $m$ i..i.d. samples of a random variable $\mathtt{R} \in [0,1]$. Let $\mu = \mathbb{E}[\mathtt{R}]$ and $\widehat{\mu} = \frac{1}{m}\sum_{i=1}^m R_i$. Fix $\varepsilon, \delta > 0$. Then, for $m \ge \frac{1}{2\varepsilon^2}\log(2/\delta)$, with probability $1-\delta$, we have that $|\mu - \widehat{\mu}| \le \varepsilon$.*

We can now use the above concentration bound in order to estimate the expected rewards. For each state-action $(\mathtt{s},\mathtt{a})$ let $\widehat{\mathtt{r}}(\mathtt{s},\mathtt{a}) = \frac{1}{m}\sum_{i=1}^m \mathtt{R}_i(\mathtt{s},\mathtt{a})$ be the average of $m$ samples. We can show the following:

**Claim 8.4.** *Given $m \ge \frac{\mathtt{R}_{\max}}{2\varepsilon^2}\log\frac{2|\mathcal{S}|\,|\mathcal{A}|}{\delta}$ samples for each state action $(\mathtt{s},\mathtt{a})$, then with probability $1-\delta$ we have for every $(\mathtt{s},\mathtt{a})$ that $|\mathtt{r}(\mathtt{s},\mathtt{a}) - \widehat{\mathtt{r}}(\mathtt{s},\mathtt{a})| \le \varepsilon$.*

*Proof.* First, we will need to scale the random variables to $[0,1]$, which will be achieved by dividing by $\mathtt{R}_{\max}$. Then, by the Chernoff-Hoffding bound (Corollary 8.3), using $\varepsilon' = \frac{\varepsilon}{\mathtt{R}_{\max}}$ and $\delta' = \frac{\delta}{|\mathcal{S}|\,|\mathcal{A}|}$, we have that for each $(\mathtt{s},\mathtt{a})$ we have that with probability $1 - \frac{\delta}{|\mathcal{S}|\,|\mathcal{A}|}$ that $|\frac{\mathtt{r}(\mathtt{s},\mathtt{a})}{\mathtt{R}_{\max}} - \frac{\widehat{\mathtt{r}}(\mathtt{s},\mathtt{a})}{\mathtt{R}_{\max}}| \le \frac{\varepsilon}{\mathtt{R}_{\max}}$.

We bound the probability over all state-action pairs using a union bound,

$$\begin{aligned}\Pr\left[\exists(\mathtt{s},\mathtt{a}) : \left|\frac{\mathtt{r}(\mathtt{s},\mathtt{a})}{\mathtt{R}_{\max}} - \frac{\widehat{\mathtt{r}}(\mathtt{s},\mathtt{a})}{\mathtt{R}_{\max}}\right| > \frac{\varepsilon}{\mathtt{R}_{\max}}\right] &\le \sum_{(\mathtt{s},\mathtt{a})} \Pr\left[\left|\frac{\mathtt{r}(\mathtt{s},\mathtt{a})}{\mathtt{R}_{\max}} - \frac{\widehat{\mathtt{r}}(\mathtt{s},\mathtt{a})}{\mathtt{R}_{\max}}\right| > \frac{\varepsilon}{\mathtt{R}_{\max}}\right] \\ &\le \sum_{(\mathtt{s},\mathtt{a})} \frac{\delta}{|\mathcal{S}|\,|\mathcal{A}|} = \delta\end{aligned}$$

Therefore, we have that with probability $1-\delta$ for every $(\mathtt{s},\mathtt{a})$ simultaneously we have $|\mathtt{r}(\mathtt{s},\mathtt{a}) - \widehat{\mathtt{r}}(\mathtt{s},\mathtt{a})| \le \varepsilon$. □

### 8.2.2 Influence of reward estimation errors

We would like to quantify the influence of having inaccurate estimates of the rewards. We will look both at the finite horizon return and the discounted return. We start with the case of finite horizon.

**Influence of reward estimation errors: Finite horizon**

Fix a deterministic Markov policy $\pi \in \Pi_{MD}$. We want to compare the return using $\mathtt{r_t}(\mathtt{s},\mathtt{a})$ versus $\widehat{\mathtt{r}}(\mathtt{s},\mathtt{a})$ and $\mathtt{r_T}(\mathtt{s})$ versus $\widehat{\mathtt{r}}_\mathtt{T}(\mathtt{s})$. We will assume that for every $(\mathtt{s},\mathtt{a})$ and $\mathtt{t}$ we have $|\mathtt{r_t}(\mathtt{s},\mathtt{a}) - \widehat{\mathtt{r}}_\mathtt{t}(\mathtt{s},\mathtt{a})| \leq \varepsilon$ and $|\mathtt{r_T}(\mathtt{s}) - \widehat{\mathtt{r}}_\mathtt{T}(\mathtt{s})| \leq \varepsilon$. We will show that the difference in return is bounded by $\varepsilon(\mathtt{T}+1)$, where $\mathtt{T}$ is the finite horizon.

Define the expected return of $\pi$ with the true rewards

$$\mathcal{V}^{\pi}_{\mathtt{T}}(\mathtt{s}_0) = \mathbb{E}^{\pi,\mathtt{s}_0}[\sum_{\mathtt{t}=0}^{\mathtt{T}} \mathtt{r_t}(\mathtt{s_t},\mathtt{a_t}) + \mathtt{r_T}(\mathtt{s_T})].$$

and with the estimated rewards

$$\widehat{\mathcal{V}}^{\pi}_{\mathtt{T}}(\mathtt{s}_0) = \mathbb{E}^{\pi,\mathtt{s}_0}[\sum_{\mathtt{t}=0}^{\mathtt{T}} \widehat{\mathtt{r}}_\mathtt{t}(\mathtt{s_t},\mathtt{a_t}) + \widehat{\mathtt{r}}_\mathtt{T}(\mathtt{s_T})].$$

We are interested in bounding the difference between the two

$$error(\pi) = |\mathcal{V}^{\pi}_{\mathtt{T}}(\mathtt{s}_0) - \widehat{\mathcal{V}}^{\pi}_{\mathtt{T}}(\mathtt{s}_0)|.$$

Note that in both cases we use the true transition probability. For a given trajectory $\sigma = (\mathtt{s}_0, \mathtt{a}_0, \dots, \mathtt{s}_{\mathtt{T}-1}, \mathtt{a}_{\mathtt{T}-1}, \mathtt{s_T})$ we define

$$error(\pi,\sigma) = \left(\sum_{\mathtt{t}=0}^{\mathtt{T}} \mathtt{r_t}(\mathtt{s_t},\mathtt{a_t}) + \mathtt{r_T}(\mathtt{s_T})\right) - \left(\sum_{\mathtt{t}=0}^{\mathtt{T}} \widehat{\mathtt{r}}_\mathtt{t}(\mathtt{s_t},\mathtt{a_t}) + \widehat{\mathtt{r}}_\mathtt{T}(\mathtt{s_T})\right)$$

taking the expectation over trajectories we have

$$error(\pi) = |\mathbb{E}^{\pi,\mathtt{s}_0}[error(\pi,\sigma)]|$$

**Lemma 8.5.** *Assume that for every* $(\mathtt{s},\mathtt{a})$ *and* $\mathtt{t}$ *we have* $|\mathtt{r_t}(\mathtt{s},\mathtt{a}) - \widehat{\mathtt{r}}_\mathtt{t}(\mathtt{s},\mathtt{a})| \leq \varepsilon$ *and for every* $\mathtt{s}$ *we have* $|\mathtt{r_T}(\mathtt{s}) - \widehat{\mathtt{r}}_\mathtt{T}(\mathtt{s})| \leq \varepsilon$*. Then, for any policy* $\pi \in \Pi_{MD}$ *we have* $error(\pi) \leq \varepsilon(\mathtt{T}+1)$*.*

*Proof.* Since $\pi \in \Pi_{MD}$ it implies that $\pi$ depends only on the time $\mathtt{t}$ and state $\mathtt{s_t}$. Therefore, probability of each trajectory $\sigma = (\mathtt{s}_0, \mathtt{a}_0, \dots, \mathtt{s}_{\mathtt{T}-1}, \mathtt{a}_{\mathtt{T}-1}, \mathtt{s_T})$ is the same under the true rewards $\mathtt{r_t}(\mathtt{s},\mathtt{a})$ and the estimated rewards $\widehat{\mathtt{r}}_\mathtt{t}(\mathtt{s},\mathtt{a})$, For each trajectory $\sigma = (\mathtt{s}_0, \mathtt{a}_0, \ldots, \mathtt{s}_{\mathtt{T}-1}, \mathtt{a}_{\mathtt{T}-1}, \mathtt{s}_{\mathtt{T}})$, we have,

$$\begin{aligned}
error(\pi, \sigma) &= \left| \sum_{\mathtt{t}=0}^{\mathtt{T}} (\mathtt{r}_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) + \mathtt{r}_\mathtt{T}(\mathtt{s}_\mathtt{T})) - \sum_{\mathtt{t}=0}^{\mathtt{T}} (\widehat{\mathtt{r}}_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) + \widehat{\mathtt{r}}_\mathtt{T}(\mathtt{s}_\mathtt{T})) \right| \\
&= \left| \sum_{\mathtt{t}=0}^{\mathtt{T}} (\mathtt{r}_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) - \widehat{\mathtt{r}}_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})) + (\mathtt{r}_\mathtt{T}(\mathtt{s}_\mathtt{T}) - \widehat{\mathtt{r}}_\mathtt{T}(\mathtt{s}_\mathtt{T})) \right| \\
&\le \sum_{\mathtt{t}=0}^{\mathtt{T}} |\mathtt{r}_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) - \widehat{\mathtt{r}}_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})| + |\mathtt{r}_\mathtt{T}(\mathtt{s}_\mathtt{T}) - \widehat{\mathtt{r}}_\mathtt{T}(\mathtt{s}_\mathtt{T})| \\
&\le \varepsilon \mathtt{T} + \varepsilon
\end{aligned}$$

The lemma follows since $error(\pi) = |E^{\pi,\mathtt{s}_0}[error(\pi, \sigma)]| \le \varepsilon(\mathtt{T}+1)$, and the bound hold for every trajectory $\sigma$. □

**Computing approximate optimal policy: finite horizon**

We now describe how to compute a near optimal policy for the finite horizon case. We start with the sample requirement. We need a sample of size $m \ge \frac{1}{2\varepsilon^2} \log \frac{2|\mathcal{S}|\,|\mathcal{A}|\,\mathtt{T}}{\delta}$ for each random variable $\mathtt{R}_\mathtt{t}(\mathtt{s}, \mathtt{a})$ and $\mathtt{R}_\mathtt{T}(\mathtt{s})$. Given the sample, we compute the rewards estimates $\widehat{\mathtt{r}}_\mathtt{t}(\mathtt{s}, \mathtt{a})$ and $\widehat{\mathtt{r}}_\mathtt{T}(\mathtt{s})$. By Claim 8.4, with probability $1-\delta$, for every $\mathtt{s} \in \mathcal{S}$ and action $\mathtt{a} \in \mathcal{A}$, we have $|\mathtt{r}_\mathtt{t}(\mathtt{s}, \mathtt{a}) - \widehat{\mathtt{r}}_((\mathtt{s}, \mathtt{a})| \le \varepsilon$ and $|\mathtt{r}_\mathtt{T}(\mathtt{s}) - \widehat{\mathtt{r}}_\mathtt{T}(\mathtt{s})| \le \varepsilon$. Now we can compute the optimal policy $\widehat{\pi}^*$ for the estimated rewards $\widehat{\mathtt{r}}_\mathtt{t}(\mathtt{s}, \mathtt{a})$ and $\widehat{\mathtt{r}}_\mathtt{T}(\mathtt{s})$. The main goal is to show that $\widehat{\pi}^*$ is a near optimal policy.

**Theorem 8.6.** *Assume that for every* $(\mathtt{s}, \mathtt{a})$ *and* $\mathtt{t}$ *we have* $|\mathtt{r}_\mathtt{t}(\mathtt{s}, \mathtt{a}) - \widehat{\mathtt{r}}_\mathtt{t}(\mathtt{s}, \mathtt{a})| \le \varepsilon$ *and for every* $\mathtt{s}$ *we have* $|\mathtt{r}_\mathtt{T}(\mathtt{s}) - \widehat{\mathtt{r}}_\mathtt{T}(\mathtt{s})| \le \varepsilon$. *Then,*

$$\mathcal{V}_\mathtt{T}^{\pi^*}(\mathtt{s}_0) - \mathcal{V}_\mathtt{T}^{\widehat{\pi}^*}(\mathtt{s}_0) \le 2\varepsilon(\mathtt{T}+1)$$

*Proof.* By Lemma 8.5, for any policy $\pi$, we have that $error(\pi) \le \varepsilon(\mathtt{T}+1)$. This implies that,

$$\mathcal{V}_\mathtt{T}^{\pi^*}(\mathtt{s}_0) - \widehat{\mathcal{V}}_\mathtt{T}^{\pi^*}(\mathtt{s}_0) \le error(\pi^*) \le \varepsilon(\mathtt{T}+1)$$

and

$$\widehat{\mathcal{V}}_\mathtt{T}^{\widehat{\pi}^*}(\mathtt{s}_0) - \mathcal{V}_\mathtt{T}^{\widehat{\pi}^*}(\mathtt{s}_0) \le error(\widehat{\pi}^*) \le \varepsilon(\mathtt{T}+1)$$

Since $\widehat{\pi}^*$ is optimal for $\widehat{\mathtt{r}}_\mathtt{t}$ we have

$$\widehat{\mathcal{V}}_\mathtt{T}^{\pi^*}(\mathtt{s}_0) \le \widehat{\mathcal{V}}_\mathtt{T}^{\widehat{\pi}^*}(\mathtt{s}_0)$$

The theorem follows by adding the three inequalities. □

**Influence of reward estimation errors: discounted return**

Fix a stationary deterministic policy $\pi \in \Pi_{SD}$. Again, define the expected return of policy $\pi$ with the true rewards

$$\mathcal{V}^{\pi}_{\gamma}(\mathtt{s}_0) = \mathbb{E}^{\pi,\mathtt{s}_0}[\sum_{\mathtt{t}=0}^{\infty} \mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})\gamma^{\mathtt{t}}]$$

and with the estimated rewards

$$\widehat{\mathcal{V}}^{\pi}_{\gamma}(\mathtt{s}_0) = \mathbb{E}^{\pi,\mathtt{s}_0}[\sum_{\mathtt{t}=0}^{\infty} \widehat{\mathtt{r}}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})\gamma^{\mathtt{t}}]$$

We are interested in bounding the difference between the two

$$error(\pi) = |\mathcal{V}^{\pi}_{\gamma}(\mathtt{s}_0) - \widehat{\mathcal{V}}^{\pi}_{\gamma}(\mathtt{s}_0)|$$

For a given trajectory $\sigma = (\mathtt{s}_0, \mathtt{a}_0, \dots, \mathtt{s}_{\mathtt{T}-1}, \mathtt{a}_{\mathtt{T}-1}, \mathtt{s}_\mathtt{T})$ we define

$$error(\pi, \sigma) = \sum_{\mathtt{t}=0}^{\infty} \gamma^{\mathtt{t}} \mathtt{r}_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) - \sum_{\mathtt{t}=0}^{\infty} \gamma^{\mathtt{t}} \hat{\mathtt{r}}_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})$$

Again, taking the expectation over trajectories we have

$$error(\pi) = |\mathbb{E}^{\pi,\mathtt{s}_0}[error(\pi, \sigma)]|$$

**Lemma 8.7.** *Assume that for every* $(\mathtt{s}, \mathtt{a})$ *we have* $|\mathtt{r}_\mathtt{t}(\mathtt{s}, \mathtt{a}) - \widehat{\mathtt{r}}_\mathtt{t}(\mathtt{s}, \mathtt{a})| \le \varepsilon$. *Then, for any policy* $\pi \in \Pi_{SD}$ *we have* $error(\pi) \le \frac{\varepsilon}{1-\gamma}$.

*Proof.* Since the policy $\pi \in \Pi_{SD}$ is stationary, the probability of each trajectory $\sigma = (\mathtt{s}_0, \mathtt{a}_0, \dots, \mathtt{s}_{\mathtt{T}-1}, \mathtt{a}_{\mathtt{T}-1}, \mathtt{s}_\mathtt{T})$ is the same under $\mathtt{r}(\mathtt{s}, \mathtt{a})$ and $\widehat{\mathtt{r}}(\mathtt{s}, \mathtt{a})$. For each trajectory $\sigma = (\mathtt{s}_0, \mathtt{a}_0, \dots, \mathtt{s}_{\mathtt{T}-1}, \mathtt{a}_{\mathtt{T}-1}, \mathtt{s}_\mathtt{T})$, we have,

$$\begin{aligned} error(\pi, \sigma) &= \left| \sum_{\mathtt{t}=0}^{\infty} \mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})\gamma^{\mathtt{t}} - \sum_{\mathtt{t}=0}^{\infty} \widehat{\mathtt{r}}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})\gamma^{\mathtt{t}} \right| \\ &= \left| \sum_{\mathtt{t}=0}^{\infty} \left(\mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) - \widehat{\mathtt{r}}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})\right) \gamma^{\mathtt{t}} \right| \\ &\le \sum_{\mathtt{t}=0}^{\infty} \left|\mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) - \widehat{\mathtt{r}}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})\right| \gamma^{\mathtt{t}} \\ &\le \frac{\varepsilon}{1-\gamma} \end{aligned}$$

The lemma follows since $error(\pi) = |E^{\pi,\mathtt{s}_0}[error(\pi, \sigma)]| \le \frac{\varepsilon}{1-\gamma}$. $\square$

**Computing approximate optimal policy: discounted return**

We now describe how to compute a near optimal policy for the discounted return. We need a sample of size $m \geq \frac{\mathtt{R}_{\max}}{2\varepsilon^2} \log \frac{2|\mathcal{S}|\,|\mathcal{A}|}{\delta}$ for each random variable $\mathtt{R}(\mathtt{s},\mathtt{a})$. Given the sample, we compute $\widehat{\mathtt{r}}(\mathtt{s},\mathtt{a})$. As we saw in the finite horizon case, with probability $1-\delta$, we have for every $(\mathtt{s},\mathtt{a})$ that $|\mathtt{r}(\mathtt{s},\mathtt{a}) - \widehat{\mathtt{r}}(\mathtt{s},\mathtt{a})| \leq \varepsilon$. Now we can compute the policy $\widehat{\pi}^*$ for the estimated rewards $\widehat{\mathtt{r}}_{\mathtt{t}}(\mathtt{s},\mathtt{a})$. Again, the main goal is to show that $\widehat{\pi}^*$ is a near optimal policy.

**Theorem 8.8.** *Assume that for every* $(\mathtt{s},\mathtt{a})$ *we have* $|\mathtt{r}(\mathtt{s},\mathtt{a}) - \widehat{\mathtt{r}}(\mathtt{s},\mathtt{a})| \leq \varepsilon$. *Then,*

$$\mathcal{V}_\gamma^{\pi^*}(\mathtt{s}_0) - \mathcal{V}_\gamma^{\widehat{\pi}^*} \leq \frac{2\varepsilon}{1-\gamma}$$

*Proof.* By Lemma 8.7 for any $\pi \in \Pi_{SD}$ we have $error(\pi) \leq \frac{\varepsilon}{1-\gamma}$. Therefore,

$$\mathcal{V}_\gamma^{\pi^*}(\mathtt{s}_0) - \widehat{\mathcal{V}}_\gamma^{\pi^*}(\mathtt{s}_0) \leq error(\pi^*) \leq \frac{\varepsilon}{1-\gamma}$$

and

$$\widehat{\mathcal{V}}_\gamma^{\widehat{\pi}^*}(\mathtt{s}_0) - \mathcal{V}_\gamma^{\widehat{\pi}^*}(\mathtt{s}_0) \leq error(\widehat{\pi}^*) \leq \frac{\varepsilon}{1-\gamma}$$

Since $\widehat{\pi}^*$ is optimal for $\widehat{\mathtt{r}}$ we have

$$\widehat{\mathcal{V}}_\gamma^{\pi^*}(\mathtt{s}_0) \leq \widehat{\mathcal{V}}_\gamma^{\widehat{\pi}^*}(\mathtt{s}_0)$$

The theroem follows by adding the three inequalities. □

### 8.2.3 Estimating the transition probabilities

We now estimate the transition probabilities. Again, we will look at the observed model. Namely, for a given state-action $(\mathtt{s},\mathtt{a})$, we consider $m$ i.i.d. transitions $(\mathtt{s},\mathtt{a},\mathtt{s}'_i)$, for $1 \leq i \leq m$. We define the observed transition distribution,

$$\widehat{p}(\mathtt{s}'|\mathtt{s},\mathtt{a}) = \frac{|\{i : \mathtt{s}'_i = \mathtt{s}'\}|}{m}$$

Our main goal would be to evaluate the observed model as a function of the sample size $m$.

We start with a general well-known observation about distributions.

**Theorem 8.9.** *Let $q_1$ and $q_2$ be two distributions over $\mathcal{S}$. Let $f : \mathcal{S} \to [0, F_{max}]$. Then,*

$$|\mathbb{E}_{s\sim q_1}[f(\mathtt{s})] - \mathbb{E}_{s\sim q_2}[f(\mathtt{s})]| \le F_{max}\|q_1 - q_2\|_1$$

*where* $\|q_1 - q_2\|_1 = \sum_{\mathtt{s}\in\mathcal{S}} |q_1(\mathtt{s}) - q_2(\mathtt{s})|$.

*Proof.* Consider the following derivation,

$$\begin{aligned}|\mathbb{E}_{s\sim q_1}[f(\mathtt{s})] - \mathbb{E}_{s\sim q_2}[f(\mathtt{s})]| &= |\sum_{\mathtt{s}\in\mathcal{S}} f(\mathtt{s})q_1(\mathtt{s}) - \sum_{\mathtt{s}\in\mathcal{S}} f(\mathtt{s})q_2(\mathtt{s})| \\ &\le \sum_{\mathtt{s}\in\mathcal{S}} f(\mathtt{s})|q_1(\mathtt{s}) - q_2(\mathtt{s})| \\ &\le F_{max}\|q_1 - q_2\|_1\end{aligned}$$

□

When we measure the distance between two Markov chains $M_1$ and $M_2$, it is natural to consider the next state distributions of each state $i$, namely $M[i,\cdot]$. The distribution for state $i$ can be measured by by the $L_1$ norm, i.e., $\|M_1[i,\cdot] - M_2[i,\cdot]\|_1$. We would like to take the worse case over states, and define $\|M\|_{\infty,1} = \max_i \sum_j |M[i,j]|$. The measure that we will consider is $\|M_1 - M_2\|_{\infty,1}$, and assume that $\|M_1 - M_2\|_{\infty,1} \le \alpha$, namely, that for any state, the next state distributions differ by at most $\alpha$ in norm $L_1$.

Clearly if $\alpha \approx 0$ then the distributions will be almost identical, but we like to get a quantitative bound on the difference, which will allow us to derive an upper bound of the required the sample size $m$.

**Theorem 8.10.** *Assume that $\|M_1 - M_2\|_{\infty,1} \le \alpha$. Let $q_1^{\mathtt{t}}$ and $q_2^{\mathtt{t}}$ be the distribution over states after trajectories of length $\mathtt{t}$ of $M_1$ and $M_2$, respectively. Then,*

$$\|q_1^{\mathtt{t}} - q_2^{\mathtt{t}}\|_1 \le \alpha\mathtt{t}$$

*Proof.* Let $p_0$ be the distribution of the start state. Then $q_1^{\mathtt{t}} = p_0^\top M_1^{\mathtt{t}}$ and $q_2^{\mathtt{t}} = p_0^\top M_2^{\mathtt{t}}$. The proof is by induction on $\mathtt{t}$. Clearly, for $\mathtt{t} = 0$ we have $q_1^0 = q_2^0 = p_0^\top$.

We start with a few basic facts about matrix norms. Recall that $\|M\|_{\infty,1} = \max_i \sum_j |M[i,j]|$. Then,

$$\|zM\|_1 = \sum_j |\sum_i z[i]M[i,j]| \le \sum_{i,j} |z[i]|\,|M[i,j]| \le \sum_i |z[i]| \sum_j |M[i,j| \le \|z\|_1\|M\|_{\infty,1} \tag{8.1}$$

This implies the following two simple facts. First, let $q$ be a distribution, i.e., $\|q\|_1 = 1$, and $M$ a matrix such that $\|M\|_{\infty,1} \leq \alpha$. Then,

$$\|qM\|_1 \leq \|q\|_1 \|M\|_{\infty,1} \leq \alpha \tag{8.2}$$

Second, let $M$ be a row-stochastic matrix, implies that $\|M\|_{\infty,1} = 1$. Then,

$$\|zM\|_1 \leq \|z\|_1 \|M\|_{\infty,1} \leq \|z\|_1 \tag{8.3}$$

For the induction step, let $z^{\mathtt{t}} = q_1^{\mathtt{t}} - q_2^{\mathtt{t}}$. We have,

$$\begin{aligned} \|q_1^{\mathtt{t}} - q_2^{\mathtt{t}}\|_1 &= \|p_0^\top M_1^{\mathtt{t}} - p_0^\top M_2^{\mathtt{t}}\|_1 \\ &= \|q_1^{\mathtt{t}-1} M_1 - (q_1^{\mathtt{t}-1} - z^{\mathtt{t}-1}) M_2\|_1 \\ &\leq \|q_1^{\mathtt{t}-1}(M_1 - M_2)\|_1 + \|z^{\mathtt{t}-1} M_2\|_1 \\ &\leq \alpha + \alpha(\mathtt{t}-1) = \alpha \mathtt{t} \end{aligned}$$

where in the last inequality is derived as follows. For the first term we used the Eq (8.2. For the second term we used Eq (8.3 with the inductive claim. □

**Approximate model and simulation lemma**

We define an *$\alpha$-approximate model* as follows.

**Definition 8.1.** *A model $\widehat{M}$ is an $\alpha$-approximate model of $M$ if for every state-action $(\mathtt{s}, \mathtt{a})$ we have: (1) $|\widehat{\mathtt{r}}(\mathtt{s}, \mathtt{a}) - \mathtt{r}(\mathtt{s}, \mathtt{a})| \leq \alpha$ and (2) $\|\widehat{p}(\cdot|\mathtt{s}, \mathtt{a}) - p(\cdot|\mathtt{s}, \mathtt{a})\|_1 \leq \alpha$.*

The following simulation lemma, for the finite horizon case, guarantees that approximate models have similar return.

**Lemma 8.11.** *Assume that model $\widehat{M}$ is an $\alpha$-approximate model of $M$. For the finite horizon return, for any policy $\pi \in \Pi_{MD}$, we have*

$$|\mathcal{V}_{\mathtt{T}}^{\pi}(\mathtt{s}_0; M) - \mathcal{V}_{\mathtt{T}}^{\pi}(\mathtt{s}_0; \widehat{M})| \leq \varepsilon$$

*for $\alpha \leq \frac{\varepsilon}{\mathtt{R}_{\max}\mathtt{T}^2}$*

*Proof.* By Theorem 8.10 the distance between the state distributions of $M$ and $\widehat{M}$ at time $\mathtt{t}$ is bounded by $\alpha\mathtt{t}$. Since the maximum reward is $\mathtt{R}_{\max}$, by Theorem 8.9 the difference is bounded by $\sum_{\mathtt{t}=0}^{\mathtt{T}} \alpha \mathtt{t} \mathtt{R}_{\max} \leq \alpha \mathtt{T}^2 \mathtt{R}_{\max}$. For $\alpha \leq \frac{\varepsilon}{\mathtt{R}_{\max}\mathtt{T}^2}$ it implies that the difference is at most $\varepsilon$. □

We now switch to the simulation lemma, for the discounted return case, which also guarantees that approximate models have similar return.

**Lemma 8.12.** *Assume that model $\widehat{M}$ is an $\alpha$-approximate model of $M$. For the discounted return, for any policy $\pi \in \Pi_{MD}$, we have*

$$|\mathcal{V}^{\pi}_{\gamma}(\mathtt{s}_0; M) - \mathcal{V}^{\pi}_{\gamma}(\mathtt{s}_0; \widehat{M})| \leq \varepsilon$$

*for* $\alpha \leq \frac{0.5\varepsilon(1-\gamma)^2}{\mathtt{R}_{\max}\log^2(\mathtt{R}_{\max}/(\varepsilon(1-\gamma)))}$. $c > 0$.

*Proof.* We reduce the discounted setting to the finite horizon setting. By Theorem 8.1, a horizon $\mathtt{T} = \frac{1}{1-\gamma}\log\frac{\mathtt{R}_{\max}}{\varepsilon(1-\gamma)/2}$, guarantees that the error due to truncating at horizon $\mathtt{T}$ is at most $\varepsilon/2$. By Lemma 8.11, for a horizon $\mathtt{T}$ and $\alpha \leq \frac{\varepsilon/2}{\mathtt{R}_{\max}|\mathtt{T}^2}$ we get and error at most $\varepsilon/2$. By substituting the bound for $\mathtt{T}$ in the expression for $\alpha$ we derive the theorem. $\square$

## Putting it all together

We want with high probability $(1-\delta)$ to have an $\alpha$-approximate model. For this we need to bound the sample size needed to approximate a distribution in the norm $L_1$. Here, Bretagnolle Huber-Carol inequality comes handy.

**Lemma 8.13** (Bretagnolle Huber-Carol). *Let $X$ be a random variable taking values in $\{1, \ldots, k\}$, where* $\Pr[X = i] = p_i]$. *Assume we sample $X$ for $n$ times and observe the value $i$ in $\hat{n}_i$ outcomes. Then,*

$$\Pr[\sum_{i=1}^{k}\left|\frac{\hat{n}_i}{n} - p_i\right| \geq \lambda] \leq 2^k e^{-n\lambda^2/2}$$

For completeness we give the proof. (The proof can also be found at Proposition A6.6 of van der Vaart and Wellner)

*Proof.* Note that,

$$\sum_{i=1}^{k}|\frac{\hat{n}_i}{n} - p_i| = \max_{S\subset[k]}\sum_{i\in S}\frac{\hat{n}_i}{n} - p_i,$$

which follows by taking $S = \{i : \frac{\hat{n}_i}{n} \geq p_i\}$.

We can now perform a concentration bound (Chernoff-Hoeffding) for each subset $S \subset [k]$, and get that the deviation is $\lambda$ with probability at most $e^{-n\lambda^2/2}$. Using a union bound over all $2^k$ subsets $S$ we derive the lemma. $\square$

The above lemma implies that to get, with probability $1-\delta$, accuracy $\alpha$ for each $(\mathtt{s},\mathtt{a})$, it is sufficient to sample $m = O(\frac{|\mathcal{S}|+\log(|\mathcal{S}|\,|\mathcal{A}|/\delta)}{\alpha^2})$ samples for each state-action pair $(\mathtt{s},\mathtt{a})$. Plugging in the value of $\alpha$ we have, for the finite horizon, we have

$$m = O(\frac{\mathtt{R}_{\max}^2}{\varepsilon^2}\mathtt{T}^4(|\mathcal{S}| + \log(|\mathcal{S}|\,|\mathcal{A}|/\delta)))$$

and for the discounted return

$$O(\frac{\mathtt{R}_{\max}^2}{\varepsilon^2}\frac{1}{(1-\gamma)^4}(|\mathcal{S}| + \log(|\mathcal{S}|\,|\mathcal{A}|/\delta)\log^4\frac{\mathtt{R}_{\max}}{\varepsilon(1-\gamma)}))$$

Assume we have a sample of $m$ for each $(\mathtt{s},\mathtt{a})$. Then with probability $1-\delta$ we have an $\alpha$-approximate model $\widehat{M}$. We compute an optimal policy $\widehat{\pi}^*$ for $\widehat{M}$. This implies that $\widehat{\pi}^*$ is a $2\varepsilon$-optimal policy. Namely,

$$|\mathcal{V}^*(\mathtt{s}_0) - \mathcal{V}^{\widehat{\pi}^*}(\mathtt{s}_0)| \leq 2\varepsilon$$

When considering the total sample size, we need to consider all state-action pairs. For the finite horizon, the total sample size is

$$m\mathtt{T}|\mathcal{S}|\,|\mathcal{A}| = O(\frac{\mathtt{R}_{\max}^2}{\varepsilon^2}|\mathcal{S}|^2|\mathcal{A}|\mathtt{T}^5\log(|\mathcal{S}|\,|\mathcal{A}|/\delta))$$

and for the discounted return

$$m|\mathcal{S}|\,|\mathcal{A}| = O(\frac{\mathtt{R}_{\max}^2}{\varepsilon^2}|\mathcal{S}|^2|\mathcal{A}|\frac{1}{(1-\gamma)^4}\log(|\mathcal{S}|\,|\mathcal{A}|/\delta)\log^4\frac{\mathtt{R}_{\max}}{\varepsilon(1-\gamma)})$$

We can now look on the dependency of our sample complexity and its dependence on the various parameters.

1. The error scales like $\frac{\mathtt{R}_{\max}^2}{\varepsilon^2}$ which looks like the right bound, even for estimation of random variables expectations.

2. The dependency on the horizon is necessary, although it is probably not optimal. In [3] a sample bound of $O(\frac{|\mathcal{S}|^2|\mathcal{A}|\mathtt{T}^2\mathtt{R}_{\max}^2}{\varepsilon^2}\log\frac{1}{\delta})$ is given.

3. The dependency on the number on the number of states $|\mathcal{S}|$ and actions $|\mathcal{A}|$, is due to the fact that we like a very high approximation of the next state distribution. We need to approximate $|\mathcal{S}|^2|\mathcal{A}|$ parameters, so for this task the bound is reasonable. However, we will show that if we restrict the task to compute an approximate optimal policy we can reduce the sample size by a factor of approximately $|\mathcal{S}|$.

### 8.2.4 Improved sample bound: Approximate Value Iteration (AVI)

We would like to exhibit a better sample complexity, for the very interesting case of deriving an approximately optimal policy. The construction and proof would use the Value Iteration algorithm (see Chapter 5.6). Recall, that the Value Iteration algorithm works as follows. Initially, we set the values arbitrarily,

$$V_0 = \{V_0(\mathtt{s})\}_{\mathtt{s}\in\mathcal{S}}$$

In iteration $n$ we compute

$$\begin{aligned} V_{n+1}(\mathtt{s}) &= \max_{\mathtt{a}\in\mathcal{A}}\{\mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma\sum_{\mathtt{s}'\in S} p(\mathtt{s}'|\mathtt{s},\mathtt{a})V_n(\mathtt{s}')\} \\ &= \max_{\mathtt{a}\in\mathcal{A}}\{\mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma E_{\mathtt{s}'\sim p(\cdot|\mathtt{s},\mathtt{a})}[V_n(\mathtt{s}')]\} \end{aligned}$$

We showed that $\lim_{n\to\infty} V_n = V^*$, and that the convergence rate is $O(\frac{\gamma^n}{1-\gamma}\mathtt{R}_{\max})$. This implies that if we run for $N$ iterations, where $N = \frac{1}{1-\gamma}\log\frac{\mathtt{R}_{\max}}{\varepsilon(1-\gamma)}$, we have an error of at most $\varepsilon$. (See Chapter 5.6.)

We would like to approximate the Value Iteration algorithm using a sample. Namely, for each $(\mathtt{s},\mathtt{a})$ we have a sample of size $m$, i.e., $\{(\mathtt{s},\mathtt{a},\mathtt{r}_i,\mathtt{s}'_i)\}_{i\in[1,m]}$ The Approximate Value Iteration (AVI) using the sample would be,

$$\widehat{V}_{n+1}(\mathtt{s}) = \max_{\mathtt{a}\in\mathcal{A}}\left\{\widehat{\mathtt{r}}(\mathtt{s},\mathtt{a}) + \gamma\frac{1}{m}\sum_{i=1}^{m}\widehat{V}_n(\mathtt{s}'_i)\right\}$$

where $\widehat{\mathtt{r}} = \frac{1}{m}\sum_{i=1}^{m}\mathtt{r}_i$.

The intuition is that if we have a large enough sample, AVI will approximate the Value Iteration. We set $m$ such that, with probability $1-\delta$, for every $(\mathtt{s},\mathtt{a})$ and any iteration $n\in[1,N]$ we have:

$$\left|\mathbb{E}[\widehat{V}_n(\mathtt{s}')] - \frac{1}{m}\sum_{i=1}^{m}\widehat{V}_n(\mathtt{s}'_i)\right| \le \varepsilon$$

and also

$$|\widehat{\mathtt{r}}(\mathtt{s},\mathtt{a}) - \mathtt{r}(\mathtt{s},\mathtt{a})| \le \varepsilon$$

This holds for $m = O(\frac{V_{max}^2}{\varepsilon^2}\log(N|\mathcal{S}|\,|\mathcal{A}|/\delta))$, where $V_{max}$ bound the maximum value. I.e., for finite horizon $V_{max} = \mathtt{T}\,\mathtt{R}_{\max}$ and for discounted return $V_{max} = \frac{\mathtt{R}_{\max}}{1-\gamma}$.

Assume that we have, for every $\mathtt{s} \in \mathcal{S}$,

$$\left|\widehat{V}_n(\mathtt{s}) - V_n(\mathtt{s})\right| \leq \lambda$$

Then

$$\begin{aligned}
\left|\widehat{V}_{n+1}(\mathtt{s}) - V_{n+1}(\mathtt{s})\right| &= \left|\widehat{\mathtt{r}}(\mathtt{s},\mathtt{a}) + \gamma \frac{1}{m}\sum_{i=1}^{m} \widehat{V}_n(\mathtt{s}'_i) - \mathtt{r}(\mathtt{s},\mathtt{a}) - \gamma E_{\mathtt{s}'\sim p(\cdot|\mathtt{s},\mathtt{a})}[V_n(\mathtt{s}')]\right| \\
&\leq |\widehat{\mathtt{r}}(\mathtt{s},\mathtt{a}) - \mathtt{r}(\mathtt{s},\mathtt{a})| + \gamma \left|\frac{1}{m}\sum_{i=1}^{m} \widehat{V}_n(\mathtt{s}'_i) - E_{\mathtt{s}'\sim p(\cdot|\mathtt{s},\mathtt{a})}[V_n(\mathtt{s}')]\right| \\
&\leq \varepsilon + \gamma\lambda \leq \lambda
\end{aligned}$$

where the last inequality holds for $\lambda \geq \frac{\varepsilon}{1-\gamma}$.

Therefore, if we sample $m = O(\frac{1}{\varepsilon^2}\log\frac{N|\mathcal{S}|\,|\mathcal{A}|}{\delta})$, we have that with probability $1-\delta$ for every $(\mathtt{s},\mathtt{a})$ the approximation error is at most $\varepsilon$. This implies that the Approximate Value Iteration has error at most $\lambda = \frac{\varepsilon}{1-\gamma}$.

The main result is that we can run Approximate Value Iteration algorithm for $N$ iterations and approximate well the optimal value function and policy.

**Theorem 8.14.** *Given for every state-action pair a sample of size*

$$m = O(\frac{1}{\varepsilon^2}\log\frac{|\mathcal{S}|\,|\mathcal{A}|\log(\frac{\mathtt{R}_{\max}}{\varepsilon(1-\gamma)})}{(1-\gamma)\delta})$$

*Running the Approximate Value Iteration for $N = \frac{1}{1-\gamma}\log\frac{\mathtt{R}_{\max}}{\varepsilon(1-\gamma)}$ results in an $\varepsilon$-approximation of the optimal value function.*

The implicit drawback of the above theorem is that we are approximating only the optimal policy, and cannot evaluate an arbitrary policy.

## 8.3 On-Policy Learning

In the off-policy setting we where given some trajectories and used them to learn and model, and using it an approximately optimal policy. Essentially, we assumed that the trajectories are exploratory enough, in the sense that each $(\mathtt{s},\mathtt{a})$ has a sufficient number of samples. In the on-line setting it is the responsibility of the learner to perform the exploration. This would be the main challenge in this section.

We will consider two (similar) tasks. The first is to reconstruct the MDP to sufficient accuracy. Given such a reconstruction we can compute the optimal policy for it and be guaranteed that it is a near optimal policy in the true MDP. The second is to reconstruct only the parts of the MDP which have a significant influence on the optimal policy. In this case we will be able to show that in most time steps we are playing near optimal action.

### 8.3.1 Learning Deterministic Decision Process

Recall that a Deterministic Decision Process (DDP) is modeled by a directed graph, where the states are the vertices, and each action is associate with an edge. For simplicity we will assume that the graph is strongly connected, i.e., there is a directed path between any two states. (See Chapter 2.)

We will start by showing how to recover the DDP. The basic idea is rather simple. We partition the state-action pairs to `known` and `unknown`. Initially all states-action pairs are unknown. Each unknown state-action that we execute is moved to known. Each time we look for a path from the current state to some unknown state-action pair. When all the state-action pairs are known we are done. This implies that we have at most $|\mathcal{S}|\ |\mathcal{A}|$ iterations, and since the maximum length of such a path is at most $|\mathcal{S}|$, the total number of time steps would be bounded by $|\mathcal{S}|^2\ |\mathcal{A}|$.

To compute a path from the known state-action pairs to some unknown state-action pair, we reduce this task to a planning task in DDP. For each known state-action pair we define a reward of zero and the next state using the observed next state. For each unknon state-action pair we define a reward of $\mathtt{R}_{\max}$ and the next state is to stay at the same state. We can now solve for the optimal policy (infinite horizon average reward) of our model. As long as there are unobserved state-action pairs, the optimal policy will reach one of them.

**Theorem 8.15.** *For any strongly connected DDP there is a strategy $\rho$ which recovers the DDP in at most $O(|\mathcal{S}|^2|\mathcal{A}|)$*

*Proof.* We first define the explored model. Given an observation set $\{(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}, \mathtt{r}_\mathtt{t}, \mathtt{s}_{\mathtt{t}+1})\}$, we define an explored model $\tilde{M}$, where $\tilde{f}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) = \mathtt{s}_{\mathtt{t}+1}$ and $\tilde{\mathtt{r}}(\mathtt{s}, \mathtt{a}) = 0$. For $(\mathtt{s}, \mathtt{a})$ which do not appear in the observation set, we define $\tilde{f}(\mathtt{s}, \mathtt{a}) = \mathtt{s}$ and $\tilde{\mathtt{r}}(\mathtt{s}, \mathtt{a}) = \mathtt{R}_{\max}$.

We can now present the on-policy exploration algorithm. Initially set $\tilde{M}_0$ to have for every $(\mathtt{s}, \mathtt{a})$ the $\tilde{f}(\mathtt{s}, \mathtt{a}) = \mathtt{s}$ and $\tilde{\mathtt{r}}(\mathtt{s}, \mathtt{a}) = \mathtt{R}_{\max}$. Initialize $\mathtt{t} = 0$. At time $\mathtt{t}$ do the following.

1. Compute $\tilde{\pi}^*_\mathtt{t}$, the optimal policy for $\tilde{M}_\mathtt{t}$, for the infinite horizon average reward return.

2. If the return of $\tilde{\pi}^*_{\mathtt{t}}$ on $\tilde{M}_{\mathtt{t}}$ is zero, then terminate.

3. Use $\mathtt{a}_{\mathtt{t}} = \tilde{\pi}^*_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}})$.

4. Observe the reward $\mathtt{r}_{\mathtt{t}}$ and the next state $\mathtt{s}_{\mathtt{t}+1}$ and add $(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}, \mathtt{r}_{\mathtt{t}}, \mathtt{s}_{\mathtt{t}+1})$ to the observation set.

5. Modify $\tilde{M}_{\mathtt{t}}$ to $\tilde{M}_{\mathtt{t}+1}$ by setting for state $\mathtt{s}_{\mathtt{t}}$ and action $\mathtt{a}_{\mathtt{t}}$ the transition $\tilde{f}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) = \mathtt{s}_{\mathtt{t}}$ and the reward $\tilde{\mathtt{r}}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) = 0$. (Note that this will have an effect only the first time we encounter $(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}})$.)

We claim that at termination we have observed each state-action pair at least once. Otherwise, there will be state-action pairs that would have a reward of $\mathtt{R}_{\max}$ and at least one of those pairs would be reachable from the current known states. So the optimal policy would have a return of $\mathtt{R}_{\max}$ contradicting the fact that it had return of zero.

After the algorithm terminates, define the following model. Given the observations during the run of the algorithm $\{(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}, \mathtt{r}_{\mathtt{t}}, \mathtt{s}_{\mathtt{t}+1})\}$, we define the observed model $\widehat{M}$, where $\widehat{f}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) = \mathtt{s}_{\mathtt{t}+1}$ and $\widehat{\mathtt{r}}(\mathtt{s}, \mathtt{a}) = \mathtt{r}_{\mathtt{t}}$. This model is exactly the true DDP $M$ since it includes all state-action pairs, and for each it has the correct reward and next state. (We are using the fact that for a DDP multiple observations of the same state and action result in identical observations.) $\square$

The above algorithm reconstructs the model completely. We can be slightly more refine. We can define an *optimistic* model, whose return upper bounds that of the true model. We can then solve for the optimal policy in the optimistic model, and if it does not reach reach a new state-action pair (after sufficiently long time) then it has to be the true optimal policy.

We first define the optimistic observed model. Given an observation set $\{(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}, \mathtt{r}_{\mathtt{t}}, \mathtt{s}_{\mathtt{t}+1})\}$, we define an optimistic observed model $\widehat{M}$, where $\widehat{f}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) = \mathtt{s}_{\mathtt{t}+1}$ and $\widehat{\mathtt{r}}(\mathtt{s}, \mathtt{a}) = \mathtt{r}_{\mathtt{t}}$. For $(\mathtt{s}, \mathtt{a})$ which do not appear in the observation set, we define $\widehat{f}(\mathtt{s}, \mathtt{a}) = \mathtt{s}$ and $\widehat{\mathtt{r}}(\mathtt{s}, \mathtt{a}) = \mathtt{R}_{\max}$.

First, we claim that for any $\pi \in \Pi_{SD}$ the optimistic observed model $\widehat{M}$ can only increase the value compared to the true model $M$. Namely,

$$\widehat{V}^{\pi}(\mathtt{s}; \widehat{M}) \geq V^{\pi}(\mathtt{s}; M)$$

The increase holds for any trajectory, and note that once $\pi$ reaches $(\mathtt{s}, \mathtt{a})$ that was not observed, its reward will be $\mathtt{R}_{\max}$ forever. (This is since $\pi \in \Pi_{SD}$.)

We can now present the on-policy learning algorithm. Initially set $\widehat{M}_0$ to have for every $(\mathtt{s}, \mathtt{a})$ the $\widehat{f}(\mathtt{s}, \mathtt{a}) = \mathtt{s}$ and $\tilde{\mathtt{r}}(\mathtt{s}, \mathtt{a}) = \mathtt{R}_{\max}$. Initialize $\mathtt{t} = 0$. At time $\mathtt{t}$ do the following.

1. Compute $\widehat{\pi}^*_{\mathtt{t}}$, the optimal policy for $\widehat{M}_{\mathtt{t}}$ with the infinite horizon average reward.
2. Use $\mathtt{a}_{\mathtt{t}} = \widehat{\pi}^*_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}})$.
3. Observe the reward $\mathtt{r}_{\mathtt{t}}$ and the next state $\mathtt{s}_{\mathtt{t}+1}$ and add $(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}, \mathtt{r}_{\mathtt{t}}, \mathtt{s}_{\mathtt{t}+1})$ to the observation set.
4. Modify $\tilde{M}_{\mathtt{t}}$ to $\tilde{M}_{\mathtt{t}+1}$ by setting for state $\mathtt{s}_{\mathtt{t}}$ and action $\mathtt{a}_{\mathtt{t}}$ the transition $\tilde{f}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) = \mathtt{s}_{\mathtt{t}}$ and the reward $\tilde{\mathtt{r}}(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}) = \mathtt{r}_{\mathtt{t}}$. (Again, note that this will have an effect only the first time we encounter $(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}})$.)

We can now state the convergence of the algorithm to the optimal policy.

**Theorem 8.16.** *After $\mathtt{T} = |\mathcal{S}|^2|\mathcal{A}|$ time steps the policy $\widehat{\pi}^*_{\mathtt{T}}$ is optimal for the true model $M$.*

*Proof.* We first claim that the model $\widehat{M}_{\mathtt{t}}$ can change at most $|\mathcal{S}|\ |\mathcal{A}|$ times (i.e., $\widehat{M}_{\mathtt{t}} \neq \widehat{M}_{\mathtt{t}+1}$). Each time we change the observed model $\widehat{M}_{\mathtt{t}}$, we observe a new $(\mathtt{s}, \mathtt{a})$ for the first time. Since there are $|\mathcal{S}|\ |\mathcal{A}|$ such pairs, this bounds the number of changes of $\widehat{M}_{\mathtt{t}}$.

Next, we show that we either make a change in $\widehat{M}_{\mathtt{t}}$ during the next $|\mathcal{S}|$ steps or we never make any more changes. The model $M$ is deterministic, if we do not change the policy in the next $|\mathcal{S}|$ time steps, the policy $\widehat{\pi}^*_{\mathtt{T}} \in \Pi_{MD}$ reach a cycle and continue on this cycle forever. Hence, the model will never change.

We showed that the number of changes is at most $|\mathcal{S}|\ |\mathcal{A}|$, and the time between changes is at most $|\mathcal{S}|$. This implies that after time $\mathtt{T} \leq |\mathcal{S}|^2|\mathcal{A}|$ we never change.

The return of $\widehat{\pi}^*_{\mathtt{T}}$ after time $\mathtt{T}$ is identical in $\widehat{M}_{\mathtt{T}}$ and $M$, since all the edges it traverses are known. Let $\widehat{V}^*$ be its return. Assume that the policy $\pi^*$ has a strictly higher return in $M$, i.e., $V^* > \widehat{V}^*$. This implies that the return of $\pi^*$ is at least $V^* > \widehat{V}^*$ in $\widehat{M}_{\mathtt{T}}$, since the rewards in $\widehat{M}_{\mathtt{T}}$ are always at least those in $M$. This contradicts the fact that $\widehat{\pi}^*_{\mathtt{T}}$ is optimal for $\widehat{M}_{\mathtt{T}}$. □

In this section we used the infinite horizon average reward, however this is not critical. If we are interested in the finite horizon, or the discounted return, we can use them to define the optimal policy, and the claims would be almost identical.

### 8.3.2 On-policy learning MDP: $E^3$

We will now extend the techniques we developed for DDP to a general MDP. We will move from infinite horizon average reward to finite horizon, mainly for simplicity, however, the techniques presented can be applied to a variety of return criteria.

The main difference between a DDP and MDP is that in a DDP it is sufficient to have a single sample $(\mathtt{s}, \mathtt{a})$ to know both the reward and the next state. In a general MDP we need to have a larger sample of $(\mathtt{s}, \mathtt{a})$ to approximate it well. (Recall that to have an $\alpha$-approximate model it is sufficient to have from each state-action pair $m = O((|\mathcal{S}| + \log(\mathtt{T}|\mathcal{S}|\ |\mathcal{A}|/\delta)))$.) Otherwise, the algorithms would be very similar.

We start with the $E^3$ (Explicit Explore or Exploit) algorithm of [6]. The algorithm learns the MDP model by sampling each state-action pair $m$ times. The main task would be to generate those $m$ samples. (A technical point would be that some states-action pairs might have very low probability under any policy, such state-action pairs would be implicitly ignored.)

As in the DDP we will maintain an explored model. Given an observation set $\{(\mathtt{s_t}, \mathtt{a_t}, \mathtt{r_t}, \mathtt{s_{t+1}})\}$, we define a state-action $(\mathtt{s}, \mathtt{a})$ pair `known` if we have $m$ times $\mathtt{t}_i$, $1 \leq i \leq m$, where $\mathtt{s}_{\mathtt{t}_i} = \mathtt{s}$ and $\mathtt{a}_{\mathtt{t}_i} = \mathtt{a}$. Otherwise it is `unknown`. We define the observed distribution of a known $(\mathtt{s}, \mathtt{a})$ to be

$$\widehat{p}(\mathtt{s}'|\mathtt{s}, \mathtt{a}) = \frac{|\{\mathtt{t}_i : \mathtt{s}_{\mathtt{t}_i+1} = \mathtt{s}\}|}{m}$$

and the observed reward to be,

$$\widehat{r}(\mathtt{s}, \mathtt{a}) = \frac{1}{m}\sum_{i=1}^{m} \mathtt{r}_{\mathtt{t}_i}$$

We define the explored model $\tilde{M}$ as follows. We add a new state $\mathtt{s}_1$. For each known $(\mathtt{s}, \mathtt{a})$, we set the next state distribution $\tilde{p}(\cdot|\mathtt{s}, \mathtt{a})$ to be the observed distribution $\widehat{p}(\cdot|\mathtt{s}, \mathtt{a})$, and the reward to be zero, i.e., $\tilde{\mathtt{r}}(\mathtt{s}, \mathtt{a}) = 0$. For unknown $(\mathtt{s}, \mathtt{a})$, we define $\tilde{p}(\mathtt{s}, \mathtt{a}) = \mathtt{s}_1$ and $\tilde{\mathtt{r}}(\mathtt{s}, \mathtt{a}) = \mathtt{R}_{\max}$. For state $\mathtt{s}_1$ we have $\tilde{p}(\mathtt{s}_1, \mathtt{a}) = \mathtt{s}_1$ and $\tilde{\mathtt{r}}(\mathtt{s}_1, \mathtt{a}) = 0$ for any action $\mathtt{a} \in \mathcal{A}$ . Note that the expected value of any policy $\pi$ in $\tilde{M}$ is *exactly* the probability it will reach an unknown state-action pair.

We can now specify the $E^3$ (Explicit Explore or Exploit) algorithm. The algorithm has three parameters: (1) $m$, how many samples we need to change a state-action from unknown to known, (2) $\mathtt{T}$, which is the finite horizon, and (3) $\varepsilon$, the accuracy parameter.

Initially all state-action pairs are unknown and we set $\tilde{M}$ accordingly. We initialize $\mathtt{t} = 0$, and at time $\mathtt{t}$ do the following.

1. Compute $\tilde{\pi}^*_{\mathtt{t}}$, the optimal policy for $\tilde{M}$, for the finite horizon return.
2. If the expected return of $\tilde{\pi}^*_{\mathtt{t}}$ on $\tilde{M}$ is less than $\varepsilon/2$, then terminate.
3. Use $\mathtt{a_t} = \tilde{\pi}^*_{\mathtt{t}}(\mathtt{s_t})$.
4. Observe the reward $\mathtt{r_t}$ and the next state $\mathtt{s_{t+1}}$ and add $(\mathtt{s_t}, \mathtt{a_t}, \mathtt{r_t}, \mathtt{s_{t+1}})$ to the observations.
5. If $(\mathtt{s_t}, \mathtt{a_t})$ became known for the first time, update $\tilde{M}$ entries for $(\mathtt{s_t}, \mathtt{a_t})$.

At termination we define $M'$ as follows. For each known $(\mathtt{s}, \mathtt{a})$, we set the next state distribution to be the observed distribution $\widehat{p}(\cdot|\mathtt{s}, \mathtt{a})$, and the reward to be the observed reward, i.e., $\widehat{r}(\mathtt{s}, \mathtt{a})$. For unknown $(\mathtt{s}, \mathtt{a})$, we can define the rewards and next state distribution arbitrarily. For concreteness, we will use the following: $\widehat{p}(\mathtt{s}, \mathtt{a}) = \mathtt{s}$ and $\widehat{\mathtt{r}}(\mathtt{s}, \mathtt{a}) = \mathtt{R}_{\max}$.

**Theorem 8.17.** *Let $m \geq \frac{2\log(2)|\mathcal{S}|+\log(|\mathcal{S}|\ |\mathcal{A}|/\delta)}{\alpha^2}$ and $\alpha = \frac{\varepsilon/4}{\mathtt{R}_{\max}\mathtt{T}^2}$. The $E^3$ (Explicit Explore or Exploit) algorithm recovers an MDP $M'$, such that for any policy $\pi$ the expected return on $M'$ and $M$ differ by at most $\varepsilon$, i.e.,*

$$|\mathcal{V}^{\pi}_{M'}(\mathtt{s}_0) - \mathcal{V}^{\pi}_{M}(\mathtt{s}_0)| \leq \varepsilon \mathtt{T} + \varepsilon.$$

*In addition, the expected number of time steps until termination is at most $O(m\mathtt{T}|\mathcal{S}|\ |\mathcal{A}|/\varepsilon)$*

*Proof.* We set the sample size $m$ such that with probability $1-\delta$ we have that for every state $\mathtt{s}$ and action $\mathtt{a}$ we have that both the observed and true next state distribution are $\alpha$ close and the difference between the observed and true reward is at most $\alpha$. Namely, $\|p(|\mathtt{s}, \mathtt{a}) - \widehat{p}(\cdot|\mathtt{s}, \mathtt{a})\|_1 \leq \alpha$ and $|\mathtt{r}(\mathtt{s}, \mathtt{a}) - \widehat{\mathtt{r}}(\mathtt{s}, \mathtt{a})| \leq \alpha$.

Let $\tilde{M}_{\mathtt{t}}$ be the model at time $\mathtt{t}$. We define $\tilde{M'}_{\mathtt{t}}$ to be the model where we replace the observed next-state distributions with the true next state distributions. Since the two models are $\alpha$-approximate, their expected return differ by at most $\alpha \mathtt{T}^2 \mathtt{R}_{\max} \leq \varepsilon/4$.

Note that the probability of reaching an unknown state in $M$ and $\tilde{M}'_{\mathtt{t}}$ at time $\mathtt{t}$ is identical. This is since the two models agree on the known states, and once an unknown state is reach, we are done.

Assume there is a policy $\pi$ that at time $\mathtt{t}$ in the true model $M$ has a probability of at least $(3/4)\varepsilon$ to reach an unknown state.(Note that the set of known and unknown states change with $\mathtt{t}$.) Recall that this implies that $\pi$ has the same probability in $\tilde{M}'_{\mathtt{t}}$. This policy $\pi$ has a probability of at least $(1/2)\varepsilon$ to reach an unknown state in $\tilde{M}_{\mathtt{t}}$ since $\tilde{M}'_{\mathtt{t}}$ and $\tilde{M}_{\mathtt{t}}$ are $\alpha$-approximate.

Similarly, once at time $\mathtt{t}$ every policy $\pi$ in the true model $M$ a probability of at most $(1/4)\varepsilon$ to reach an unknown state, then we are guarantee to terminate. This is since the probability of $\pi$ to reach an unknown state is identical in $M$ and $\tilde{M}'_{\mathtt{t}}$. Since the expected return of $\pi$ in $\tilde{M}'_{\mathtt{t}}$ and $\tilde{M}$ differ by at most $\varepsilon/4$, the probability of $\pi$ to reach an unknown state in $\tilde{M}_{\mathtt{t}}$ is at most $\varepsilon/2$. This is exactly our termination condition. This also implies that in expectation we have $O(1/\varepsilon)$ iterations until termination.

Assume termination at time $\mathtt{t}$. At time $\mathtt{t}$ every policy $\pi$ has a probability of at most $(1/2)\varepsilon$ to reach some unknown state in $\tilde{M}_{\mathtt{t}}$. This implies that $\pi$ has a probability of at most $(3/4)\varepsilon$ to reach some unknown state in $M$.

After the algorithm terminates, we define the model $M'$ using the observed distributions and rewards for any known state-action pair. Since every known state-action pair is sampled $m$ times, we have that with probability $1-\delta$ the model $M'$ $\alpha$-approximation of the true model $M$, in the known state-action pairs.

When we compare $|\mathcal{V}^{\pi}_{M'}(\mathtt{s}_0) - \mathcal{V}^{\pi}_{M}(\mathtt{s}_0)$ we separate the difference due to unknown states and known states. The contribution of unknown states is at most $\varepsilon\mathtt{T}$, since the probability of reaching them is at most $(3/4)\varepsilon < \varepsilon$ and the maximum return is $\mathtt{T}$. The difference in the known states is at most $\varepsilon/4 < \varepsilon$ since $M$ and $M'$ are $\alpha$ approximate, and the selection of $\alpha$ guarantees that the difference in expectation is at most $\varepsilon/4$ (Lemma 8.11).

In each iteration, until we terminate, we have a probability of at least $\varepsilon/4$ to reach an unknown state-action. We can reach unknown state-action pairs at most $m|\mathcal{S}|\,|\mathcal{A}|$. Therefore the expected number of time steps is $O(m\mathtt{T}|\mathcal{S}|\,|\mathcal{A}|/\varepsilon)$. $\square$

### 8.3.3 On-policy learning MDP: `R-max`

In this section we introduce `R-max`. The main difference between `R-max` and $E^3$ is that `R-max` will have a single continuous phase, and there will be no need to explicitly switch from exploration to exploitation.

Similar to the DDP, we will use the principle of *Optimism in face of uncertainty.* Namely, we substitute the unknown quantities by the maximum possible values. In addition, Similar to DDP and $E^3$, we will partition the state-action pairs $(\mathtt{s},\mathtt{a})$ `known` and `unknown`. The main difference from DDP, and similar to $E^3$, is that in a DDP it is sufficient to have a single sample to move $(\mathtt{s},\mathtt{a})$ from `unknown` to `known`. In a general MDP we need have a larger sample to move $(\mathtt{s},\mathtt{a})$ from `unknown` to `known`. Otherwise, the `R-max` algorithm would be very similar to the one in DDP. In the following, we describe algorithm `R-max`, which performs on-policy learning of MDPs.

*Initialization:* Initially, we set for each $(\mathtt{s},\mathtt{a})$ a next state distribution which always return to $\mathtt{s}$, i.e., $p(\mathtt{s}|\mathtt{s},\mathtt{a}) = 1$ and $p(\mathtt{s}'|\mathtt{s},\mathtt{a}) = 0$ for $\mathtt{s}' \neq s$. We set the reward to be maximal, i.e., $\mathtt{r}(\mathtt{s},\mathtt{a}) = \mathtt{R}_{\max}$. We mark $(\mathtt{s},\mathtt{a})$ to be `unknown`.

*Execution:* At time $\mathtt{t}$. (1) Build a model $\widehat{M}_{\mathtt{t}}$, explained later. (2) Compute $\widehat{\pi}^*_{\mathtt{t}}$ the optimal finite horizon policy for $\widehat{M}_{\mathtt{t}}$, and (3) Execute $\mathtt{a}_{\mathtt{t}} = \widehat{\pi}^*_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}})$ and observe $\mathtt{r}_{\mathtt{t}}$ and $\mathtt{s}_{\mathtt{t}+1}$.

*Building a model:* At time $\mathtt{t}$, if the number of samples of $(\mathtt{s},\mathtt{a})$ is *exactly* $m$ for the *first time*, then: modify $p(\cdot|\mathtt{s},\mathtt{a})$ to the observed transition distribution, $\mathtt{r}(\mathtt{s},\mathtt{a})$ to the average observed reward for $(\mathtt{s},\mathtt{a})$ and mark $(\mathtt{s},\mathtt{a})$ as `known`. Note that we update each $(\mathtt{s},\mathtt{a})$ only once, when it moves from `unknown` to `known`.

Given a set of observations, we define the model $\tilde{M}$ as follows. For each known $(\mathtt{s},\mathtt{a})$, we set the next state distribution $\tilde{p}$ to be the observed distribution $\widehat{p}(\cdot|\mathtt{s},\mathtt{a})$, and the reward $\tilde{\mathtt{r}}$ to be the observed distribution, i.e., $\widehat{\mathtt{r}}(\mathtt{s},\mathtt{a})$. For unknown $(\mathtt{s},\mathtt{a})$, we define $\tilde{p}(\mathtt{s},\mathtt{a}) = \mathtt{s}$ and $\tilde{\mathtt{r}}(\mathtt{s},\mathtt{a}) = \mathtt{R}_{\max}$. Note that the main difference from $E^3$ is the fact that we set the rewards of the know state-action pairs to be their observed reward (and not zero, at $E^3$ does).

We can now specify the `R-max` algorithm. The algorithm has two parameters: (1) $m$, how many samples we need to change a state-action from unknown to known, and (2) $\mathtt{T}$, which is the finite horizon.

Initially all state-action pairs are unknown and we set $\tilde{M}$ accordingly. We initialize the time $\mathtt{t} = 0$, and at $\mathtt{t}$ do the following.

1. Let $k$ be the number of steps remaining in the current episode. Compute $\tilde{\pi}^*_{\mathtt{t}}$, the optimal policy for $\tilde{M}$, for the finite horizon return with horizon $k$.

2. Use $\mathtt{a}_{\mathtt{t}} = \tilde{\pi}^*_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}})$.

3. Observe the reward $\mathtt{r}_{\mathtt{t}}$ and the next state $\mathtt{s}_{\mathtt{t}+1}$ and add $(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}}, \mathtt{r}_{\mathtt{t}}, \mathtt{s}_{\mathtt{t}+1})$ to the observations.

4. If $(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}})$ became known for the first time, update $\tilde{M}$ entries for $(\mathtt{s}_{\mathtt{t}}, \mathtt{a}_{\mathtt{t}})$.

Note that there are two main differences from $E^3$. First, when a state-action becomes known, we set the reward to be the observed reward (and not zero, as in $E^3$). Second, there is no test for termination, but we continuously run the same algorithm (although at some point the policy will stop changing).

Here is the basic intuition for algorithm `R-max`. We consider the finite horizon return with horizon $\mathtt{T}$. In each episode we run $\widehat{\pi}^*_{\mathtt{t}}$ for $\mathtt{T}$ time steps. Either, with some non-negligible probability we explore a state-action $(\mathtt{s},\mathtt{a})$ which is `unknown`, in this case we make progress on the exploration. This can happen at most $m|\mathcal{S}|\,|\mathcal{A}|$ times. Alternatively, with high probability we do not reach any state-action $(\mathtt{s},\mathtt{a})$ which is `unknown`, in which case we are optimal on the observed model, and near optimal on the true model.

For the analysis define an event $NEW$, which is that event that we visit some `unknown` state-action $(\mathtt{s}, \mathtt{a})$ during the episode (of $\mathtt{T}$ time steps). For the return of $\widehat{\pi}^*_{\mathtt{t}}$, when $\mathtt{t}$ is a start of an episode (i.e., $k = \mathtt{T}$), we have,

$$V^{\widehat{\pi}^*_{\mathtt{t}}}(\mathtt{s}_0) \geq V^*(\mathtt{s}_0) - \Pr[W]\mathtt{V}_{\max} - \lambda$$

where $\mathtt{V}_{\max} = \mathtt{T}$ is the maximum $\mathtt{T}$ return, and $\lambda$ is the approximation error, and we can bound it by $\varepsilon/2$ by setting $m$ large enough.

We consider two cases, depending on the probability of $NEW$. First, we consider the case that the probability of $NEW$ is small. If $\Pr[NEW] \leq \frac{\varepsilon}{2\mathtt{V}_{\max}}$, then $V^{\widehat{\pi}^*_{\mathtt{t}}}(\mathtt{s}_0) \geq V^*(\mathtt{s}_0) - \varepsilon/2 - \varepsilon/2$, since we assume that $\lambda \leq \varepsilon/2$.

Second, we consider the case that the probability of $NEW$ is large. If $\Pr[NEW] > \frac{\varepsilon}{2\mathtt{V}_{\max}}$. Then, there is a good probability to visit an `unknown` $(\mathtt{s}, \mathtt{a})$, but this can happen at most $m|\mathcal{S}|\,|\mathcal{A}|$. Therefore, the expected number of such blocks is at most $m|\mathcal{S}|\,|\mathcal{A}|\frac{2\mathtt{V}_{\max}}{\varepsilon}$.

This implies that only in

$$m|\mathcal{S}|\,|\mathcal{A}|\frac{2\mathtt{V}_{\max}}{\varepsilon}$$

episodes, the algorithm `R-MAX` will be more than $\varepsilon$ sub-optimal, i.e., have an expected return less than $V^* - \varepsilon$.

## 8.4 Bibliography Remarks

The first polynomial time model based learning algorithm is $E^3$ (Explicit Explore or Exploit) of [6]. While we did not outline the $E^3$ algorithm, we did describe the effective horizon and the simulation lemma from there.

The improved sampling bounds for the optimal policy using approximate Value Iteration is following [5].

The `R-MAX` algorithm is due to [2].

Analysis of related models, especially the `PAC-MDP` model appears in [13, 8].

# Chapter 9

# Tabular learning: Model free

In this chapter we consider model-free learning algorithms. The main idea of model-free algorithms is to avoid learning the MDP model directly. The model based methodology was the following. During the learning we estimate model of the MDP, and later, we derive the optimal policy of the estimated model. The main point was that an optimal policy of a near-accurate MDP is an near-optimal policy in the true MDP.

The model-free methodology is going to be different. We will never learn an estimated model, but rather we will directly learn the value function of the MDP. The value function can be either the $Q$-function (as is the case in Q-learning and SARSA) or the $V$-function (as is the case in Temporal Difference (TD) algorithms and the Monte-Carlo approach).

We will first look at the Q-learning algorithms, and its on-policy variant SARSA. Then, we will look at Monte-Carlo methods. We will continue with temporal differences algorithms, $TD(\lambda)$. At the end of the chapter we have a few miscellaneous topics, including, evaluating one policy while following a different policy (using importance sampling) and the actor-critic methodology.

## 9.1 Online approximation of mean

Before we start looking at model-free learning algorithms, we look at a very simple task, approximating the average of a random variable using samples. Our main goal would be to do it in an online way while maintaining minimal information, namely the average.

Assume we have a random variable $Z \in [0,1]$ with $\mu = E[Z]$. We observe $m$ samples, $z_1, \ldots, z_m$. In the batch setting we simply compute the average of the samples, $\widehat{\mu}_m = (1/m)\sum_{i=1}^m z_i$. We can bound $|\mu - \widehat{\mu}_T|$ using Chernoff-Hoeffding concentration bounds (see Lemma 8.2).

We can actually perform the computation of the average in an online way.

$$\widehat{\mu}_m = \frac{1}{m}\sum_{i=1}^m z_i = \frac{m-1}{m}\widehat{\mu}_{m-1} + \frac{1}{m}z_m = \widehat{\mu}_{m-1} + \frac{1}{m}(z_m - \widehat{\mu}_{m-1})$$

This lead naturally to the *exponential averaging* where we have a sequence of $\alpha_t \in (0,1)$ and

$$\bar{\mu}_m = \bar{\mu}_{m-1} + \alpha_m(z_m - \bar{\mu}_{m-1}) = \sum_{i=1}^m \beta_i z_i$$

where $\beta_i = \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j)$. Note that $\sum_{i=1}^T \beta_i = 1$. The regular averaging is a special case with $\alpha_t = 1/t$.

We would like to show that the approximation $\bar{\mu}$ concentrates around the mean $\mu$. For this we will introduce the McDiarmid inequality.

The setting of the McDiarmid inequality is the following. There is a domain $\mathcal{X}$, which can be for example $\mathbb{R}^d$. We have $n$ random variables $X_i \in \mathcal{X}$. There is a function $f$ which maps the realization of the $n$ independent (but not necessarily identical) random variables to a value. Namely, $f : \mathcal{X}^n \to \mathbb{R}$.

We define $c_i$, the sensitivity of the function $f$ to the $i$-th input, to be

$$c_i = \max_{x\in\mathcal{X}^n} \max_{a_1,a_0\in X} |f(x_1,\ldots,x_{i-1},a_0,x_{i+1},\ldots,x_n) - f(x_1,\ldots,x_{i-1},a_1,x_{i+1},\ldots,x_n)|$$

The basic idea of McDimid's inequality is that with high probability the observed value of $f$ is close to its expected value. Formally,

**Lemma 9.1** (McDiarmid's inequality)**.**

$$\Pr[|f(x) - E[f(x)| \geq \varepsilon] \leq e^{-2\varepsilon^2/(\sum_i c_i^2)}$$

We can use McDirmid's inequality to recover the concentration bounds for a simple average. Define $avg(x_1,\ldots,x_n) = (1/n)\sum_{i=1}^n x_i$, where $x_i \in [0,1]$. The sensitivity of $avg$ to any input $i$ is exactly $c_i = 1/n$. This implies that $\sum_i c_i^2 = \sum_i 1/n^2 = 1/n$. Therefore, McDiarmid's inequality gives us,

$$\Pr[|avg(x) - E[avg(x)]| \geq \varepsilon] \leq e^{-2\varepsilon^2 n}$$

Note that the bound is very similar to the Chernoff-Hoeffding bound (Lemma 8.2).

We can now use the McDiarmid's inequality for the weighted average case. Define $wavg(x_1, \ldots, x_n) = \sum_{i=1}^n \beta_i x_i$, where $x_i \in [0,1]$ and $\sum_i \beta_i = 1$. The sensitivity of $wavg$ to any input $i$ is exactly $c_i = \beta_i$. Therefore, McDiarmid's inequality gives us,

$$\Pr[|wavg(x) - E[wavg(x)]| \geq \varepsilon] \leq e^{-\varepsilon^2/(\sum_i \beta_i^2)}$$

For the case of exponential averaging, with a fixed parameter $\alpha$, we have $\beta_i = \alpha(1-\alpha)^{i-1}$. This implies that

$$\sum_{i=1}^{m} \beta_i^2 = \alpha^2 \frac{1-(1-\alpha)^m}{1-(1-\alpha)^2} \approx \frac{\alpha}{2-\alpha}$$

Therefore, if we set $\alpha \approx \varepsilon^2/(\log(1/\delta))$ we have an accuracy $\varepsilon$ with confidence $1-\delta$.

## 9.2 $Q$-Learning

### 9.2.1 $Q$-Learning: Deterministic Decision Process

To demonstrate some key ideas of $Q$-learning, we start with a simplified learning algorithm that is suitable for a *Deterministic Decision Process (DDP)* model, namely:

$$\begin{aligned} \mathtt{s}_{\mathtt{t}+1} &= f(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) \\ \mathtt{r}_\mathtt{t} &= \mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) \end{aligned}$$

We consider the discounted return criterion:

$$\begin{aligned} \mathcal{V}^\pi(\mathtt{s}) &= \sum_{\mathtt{t}=0}^{\infty} \gamma^\mathtt{t} \mathtt{r}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}), \quad \text{given } \mathtt{s}_0 = \mathtt{s}, \mathtt{a}_\mathtt{t} = \pi(\mathtt{s}_\mathtt{t}) \\ \mathcal{V}^*(\mathtt{s}) &= \max_\pi \mathcal{V}^\pi(\mathtt{s}) \end{aligned}$$

where $\mathcal{V}^*$ is the value function of the optimal policy.

Recall our definition of the $Q$-function (or *state-action value function*), specialized to the present deterministic setting:

$$Q^*(\mathtt{s}, \mathtt{a}) = \mathtt{r}(\mathtt{s}, \mathtt{a}) + \gamma \mathcal{V}^*(f(\mathtt{s}, \mathtt{a}))$$

The optimality equation is then

$$\mathcal{V}^*(\mathtt{s}) = \max_\mathtt{a} Q^*(\mathtt{s}, \mathtt{a})$$

or, in terms of $Q^*$ only:

$$Q^*(\mathtt{s},\mathtt{a}) = \mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \max_{\mathtt{a}'} Q^*(f(\mathtt{s},\mathtt{a}),\mathtt{a}')$$

The $Q$ learning algorithm runs as follows:

- *Initialize:* Set $\widehat{Q}(\mathtt{s},\mathtt{a}) = 0$, for all $\mathtt{s}$, $\mathtt{a}$.
- At time $\mathtt{t} = 0, 1, \ldots$ :
  - Observe $(\mathtt{s_t},\ \mathtt{a_t},\ \mathtt{r_t},\ \mathtt{s}'_\mathtt{t})$, where $\mathtt{r_t}$ is the observed reward when executing action $\mathtt{a_t}$ in state $\mathtt{s_t}$ and $\mathtt{s}'_\mathtt{t}$ is the next state, i.e., $\mathtt{s}'_\mathtt{t} = f(\mathtt{s_t},\mathtt{a_t})$.
  - Update: $\widehat{Q}_{\mathtt{t}+1}(\mathtt{s_t},\mathtt{a_t}) := \mathtt{r_t} + \gamma \max_{\mathtt{a}'} \widehat{Q}_\mathtt{t}(\mathtt{s}'_\mathtt{t},\mathtt{a}')$

The Q-learning algorithm is an off-policy algorithm, namely, it does not select actions, but only observes actions selected by a potentially different algorithm. Specifically, note that the algorithm does not specify how to choose the actions $\mathtt{a_t}$. We will need to have some assumption about the sequence of actions selected, as is the case in the theorem below.

**Theorem 9.2** (Convergence of $Q$-learning for DDP)**.**
*Assume a DDP model. If each state-action pair is visited* infinitely-often*, then* $\lim_{\mathtt{t}\to\infty} \widehat{Q}_\mathtt{t}(\mathtt{s},\mathtt{a}) = Q^*(\mathtt{s},\mathtt{a})$*, for all* $(\mathtt{s},\mathtt{a})$.

*Proof.* Let

$$\Delta_\mathtt{t} \triangleq \|\widehat{Q}_\mathtt{t} - Q^*\|_\infty = \max_{\mathtt{s},\mathtt{a}} |\widehat{Q}_\mathtt{t}(\mathtt{s},\mathtt{a}) - Q^*(\mathtt{s},\mathtt{a})|\,.$$

Then, at every stage $\mathtt{t}$:

$$\begin{aligned}|\widehat{Q}_{\mathtt{t}+1}(\mathtt{s_t},\mathtt{a_t}) - Q^*(\mathtt{s_t},\mathtt{a_t})| &= \left|(\mathtt{r_t} + \gamma \max_{\mathtt{a}'} \widehat{Q}_\mathtt{t}(\mathtt{s}'_\mathtt{t},\mathtt{a}')) - (\mathtt{r_t} + \gamma \max_{\mathtt{a}''} Q^*(\mathtt{s}'_\mathtt{t},\mathtt{a}''))\right| \\ &= \gamma |\max_{\mathtt{a}'} \widehat{Q}_\mathtt{t}(\mathtt{s}'_\mathtt{t},\mathtt{a}') - \max_{\mathtt{a}''} Q^*(\mathtt{s}'_\mathtt{t},\mathtt{a}'')| \\ &\le \gamma \max_{\mathtt{a}'} |\widehat{Q}_\mathtt{t}(\mathtt{s}'_\mathtt{t},\mathtt{a}') - Q^*(\mathtt{s}'_\mathtt{t},\mathtt{a}')| \ \le\ \gamma\Delta_\mathtt{t}\,.\end{aligned}$$

Consider now some interval $[\mathtt{t},\mathtt{t}_1]$ over which all state-action pairs $(\mathtt{s},\mathtt{a})$ appear at least once. Using the above relation and simple induction, it follows that $\Delta_{\mathtt{t}_1} \le \gamma\Delta_\mathtt{t}$. Since each state-action pair is visited infinitely often, there is an infinite number of such intervals, and since $\gamma < 1$, it follows that $\Delta_\mathtt{t} \to 0$, as $\mathtt{t}$ goes to infinity. □

*Remark:* Note that the $Q$ learning algorithm does not need to receive a continuous trajectory, but can receive arbitrary quadruples $(\mathtt{s_t},\mathtt{a_t},\mathtt{r_t},\mathtt{s}'_\mathtt{t})$,

### 9.2.2 Q-learning: Markov Decision Process

We now extend the $Q$ learning algorithm from DDP to MDP. The main difference is that now we will need to average multiple observations to converge to the value of $Q*$. For this we introduce generate for learning rates, $\alpha_{\mathtt{t}}(\mathtt{s},\mathtt{a})$. We allow the learning rate to depend both on the state $\mathtt{s}$, action $\mathtt{a}$ and time $\mathtt{t}$. For example $\alpha_{\mathtt{t}}(\mathtt{s},\mathtt{a}) = 1/n$ where $n$ is the number of times we updated $(\mathtt{s},\mathtt{a})$ up to time $\mathtt{t}$. The following is the definition of the algorithm.

**The $Q$-learning algorithm:**

- *Initialize:* Set $Q_0(\mathtt{s},\mathtt{a}) = 0$, for all $\mathtt{s}$, $\mathtt{a}$.
- At time $\mathtt{t} = 0, 1, \ldots$:
  - Observe: $(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}, \mathtt{r}_\mathtt{t}, \mathtt{s}'_\mathtt{t})$.
  - Update:

$$Q_{\mathtt{t}+1}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t}) := Q_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t}) + \alpha_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})[\mathtt{r}_\mathtt{t} + \gamma \max_{\mathtt{a}'} Q_\mathtt{t}(\mathtt{s}'_\mathtt{t},\mathtt{a}') - Q_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})]$$

It is worth to try and gain some intuition regarding the $Q$ learning algorithm. Let $\Gamma_\mathtt{t} = \mathtt{r}_\mathtt{t} + \gamma \max_{\mathtt{a}'} Q_\mathtt{t}(\mathtt{s}'_\mathtt{t},\mathtt{a}') - Q_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})$. For simplicity assume we already converged, $Q_\mathtt{t} = Q^*$. Then we have that $E[\Gamma_\mathtt{t}] = 0$ and (on average) we maintain that $Q_\mathtt{t} = Q^*$. Clearly we do not want to assume that we converge, since this is the entire goal of the algorithm. The main challenge in showing the convergence is that in the updates we use $Q_\mathtt{t}$ rather than $Q^*$. We also need to handle the stochastic nature of the updates, where there are both stochastic rewards and stochastic next state.

The next theorem states the main convergence property of $Q$-leanring.

**Theorem 9.3** ($Q$-learning convergence)**.**
*Assume every state-action pair* $(\mathtt{s},\mathtt{a})$ *occurs infinitely often, and the step size* $\alpha_\mathtt{t}(\mathtt{s},\mathtt{a})$ *has the properties: (1)* $\sum_\mathtt{t} \alpha_\mathtt{t}(\mathtt{s},\mathtt{a}) I(\mathtt{s}_\mathtt{t} = \mathtt{s}, \mathtt{a}_\mathtt{t} = \mathtt{a}) = \infty$*, and (2)* $\sum_\mathtt{t} \alpha^2_\mathtt{t}(\mathtt{s},\mathtt{a}) I(\mathtt{s}_\mathtt{t} = \mathtt{s}, \mathtt{a}_\mathtt{t} = \mathtt{a}) = O(1)$*,*
*Then,* $Q_\mathtt{t}$ *converges with probability* $1$ *to* $Q^*$

Note that the statement of the theorem has two requirements. The first is that every state-action pair is executed infinitely often. This is clearly required for convergence (per state-action). Since $Q$ learning is an off-policy, it has no influence on the sequence of state-action it observes, and therefore we have to make this assumption. The second requirement is two properties regarding the learning rates $\alpha$. The first states that the learning rates are large enough that we can (potentially) reach an value. The second states that the learning rates are sufficiently small (sum of squares finite) so that we will be able to converge locally.

We will show the converges proof by introducing a general technique of stochastic approximation. (We will not give the details of the proof of stochastic approximation, but will only outline the main building blocks and the methodology.) Given the stochastic approximation convergence theorem, we later show how to use it to derive he convergence of the $Q$-learning algorithm.

### 9.2.3 Stochastic Approximation

There is a general framework of stochastic approximation algorithms. We will outline the main definitions and results of that literature. We later use it to show convergence of online learning algorithms.

The stochastic approximation has a state space $\mathcal{S}$, and updates an approximation $X(\mathtt{s})$ at each iteration. The updates are in the direction of $(HX)(\mathtt{s}) + w$ where $H$ is a (possibly non-linear) operator and $\omega$ is a zero mean bounded noise. The iterative algorithm takes the following general form:

$$X_{\mathtt{t}+1}(\mathtt{s}) = (1 - \alpha_{\mathtt{t}}(\mathtt{s}))X_{\mathtt{t}}(\mathtt{s}) + \alpha_{\mathtt{t}}(\mathtt{s})((HX_{\mathtt{t}})(\mathtt{s}) + \omega_{\mathtt{t}}(\mathtt{s}))$$

We will mainly look at $(B, \gamma)$ *well behaved* iterative algorithms, where $B > 0$ and $\gamma \in (0, 1)$, which have the following properties:

1. Step size: For every $(\mathtt{s}, \mathtt{a})$ we have (1) $\sum_{\mathtt{t}} \alpha_{\mathtt{t}}(\mathtt{s}, \mathtt{a}) I(\mathtt{s}_{\mathtt{t}} = \mathtt{s}, \mathtt{a}_{\mathtt{t}} = \mathtt{a}) = \infty$, and (2) $\sum_{\mathtt{t}} \alpha^2_{\mathtt{t}}(\mathtt{s}, \mathtt{a}) I(\mathtt{s}_{\mathtt{t}} = \mathtt{s}, \mathtt{a}_{\mathtt{t}} = \mathtt{a}) = O(1)$,

2. Noise: $E[\omega_{\mathtt{t}}(\mathtt{s}) | h_{\mathtt{t}-1}] = 0$ and $|\omega_{\mathtt{t}}(\mathtt{s})| \leq B$, where $h_{\mathtt{t}-1}$ is the history up to time $\mathtt{t}$.

3. Contraction: There exists $X^*$ such that for any $X$ we have $\|HX - X^*\|_\infty \leq \gamma \|X - X^*\|_\infty$. (This implies that $HX^* = X^*$. Note that since $H$ is a contracting operator, this property is guarantee to holds.)

The following is the convergence theorem for well behaved iterative algorithms.

**Theorem 9.4** (Iterative Stochastic Approximation: convergence)**.**
*Let $X_{\mathtt{t}}$ be a sequence that is generated by a $(B, \gamma)$ well behaved iterative algorithm. Then $X_{\mathtt{t}}$ converges with probability 1 to $X^*$.*

We will not give a proof of this important theorem, but we will try to sketch the main proof methodology.

There are two distinct parts to the iterative algorithms. The part $(HX_{\mathtt{t}})$ is contracting, in a deterministic manner. If we had only this part (say, $\omega_{\mathtt{t}} = 0$ always) then the contraction property of $H$ will give the convergence (as we saw before). The main challenge is the addition stochastic noise $\omega_{\mathtt{t}}$. The noise is unbiased, so on average the expectation is zero. Also, the noise is bounded by a constant $B$. This implies that if we average the noise over a long time interval, then the average should be very close to zero.

The proof considers the $\|X_{\mathtt{t}} - X^*\|$, and works in phases. In phase $i$, at any time $\mathtt{t}$ in the phase we have $\|X_{\mathtt{t}} - X^*\| \leq \lambda_i$. In each phase we have a deterministic contraction using the operator $H$. This implies that the deterministic contraction implies that the space contracts by a factor $\gamma < 1$. We have to take care of the stochastic noise. We make the phase long enough so that the average of the noise is less than $(1-\gamma)/2$ factor. This implies that the space contracts by $\lambda_{i+1} \leq \lambda_i(1+\gamma)/2 < \lambda_i$. This implies that as the number of phases increases we that that $\lambda_i < ((1+\gamma)/2)^i$ converges to zero.

### 9.2.4 $Q$-learning as a stochastic approximation

After we introduced the stochastic approximation algorithms, and their convergence theorem, we show that the $Q$ learning algorithm is a stochastic approximation algorithm, and thus converges. To show that $Q$ learning algorithm is a stochastic approximation algorithm we need to introduce an operator $H$ and the noise $\omega$.

We first define the operator $H$.

$$(Hq)(\mathtt{s},\mathtt{a}) = \sum_{\mathtt{s}'} p(\mathtt{s}'|\mathtt{s},\mathtt{a})[\mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \max_{\mathtt{a}'} q(\mathtt{s}',\mathtt{a}')]$$

The contraction of $H$ is established as follows,

$$\begin{aligned}
\|Hq_1 - Hq_2\|_\infty &= \gamma \max_{\mathtt{s},\mathtt{a}} |\sum_{\mathtt{s}'} p(\mathtt{s}'|\mathtt{s},\mathtt{a})[\max_{b_1} q_1(\mathtt{s}',b_1) - \max_{b_2} q_2(\mathtt{s}',b_2)| \\
&\leq \gamma \max_{\mathtt{s},\mathtt{a}} \max_{b,\mathtt{s}'} |q_1(\mathtt{s}',b) - q_2(\mathtt{s}',b)| \\
&\leq \gamma \|q_1 - q_2\|_\infty
\end{aligned}$$

In this section we re-write the $Q$-learning algorithm to follow the iterative stochastic approximation algorithms, so that we will be able to apply Theorem 9.4.

Recall that

$$Q_{\mathtt{t}+1}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t}) := (1-\alpha_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t}))Q_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t}) + \alpha_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})[\mathtt{r}_\mathtt{t} + \gamma \max_{\mathtt{a}'} Q_\mathtt{t}(\mathtt{s}'_\mathtt{t},\mathtt{a}') - Q_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})]$$

Let $\Phi_\mathtt{t} = \mathtt{r}_\mathtt{t} + \gamma \max_{\mathtt{a}'} Q_\mathtt{t}(\mathtt{s}_{\mathtt{t}+1},\mathtt{a}')$. This implies that $E[\Phi_\mathtt{t}] = (HQ_\mathtt{t})(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})$. We can define the noise term as $\omega_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t}) = \Phi_\mathtt{t} - (HQ_\mathtt{t})(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})$ and have $E[\omega_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})] = 0$. In addition $|\omega_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})| \le \mathtt{V}_{\max} = \frac{R_{max}}{1-\gamma}$.

We can now rewrite the $Q$-learning, as follows,

$$Q_{\mathtt{t}+1}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t}) := (1-\alpha_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t}))Q_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t}) + \alpha_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})[(HQ_\mathtt{t})(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t}) + \omega_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})]$$

In order to apply Theorem 9.4, we have the properties on the noise $\omega_\mathtt{t}$, and of the contraction of $H$. Therefore, we can derive Theorem 9.3, since the step size requirement is part of the theorem.

### 9.2.5 Step size

The step size has important implication for the convergence of the Q-learning algorithm, and more importantly, on the rate convergence. For the convergence, we need for the step size to have two properties. The first is that sum diverges, i.e., $\sum_\mathtt{t} \alpha_\mathtt{t}(\mathtt{s},\mathtt{a}) I(\mathtt{s}_\mathtt{t} = \mathtt{s}, \mathtt{a}_\mathtt{t} = \mathtt{a}) = \infty$, which intuitively implies that we can potentially reach any value. This is important, since we might have errors on the way, and this guarantees that the step sizes are large enough to possibly reach correct any error. (It does not guarantee that it will correct the errors, only that the step size is large enough to allow for it.)

The second requirement from the step size is that the sum of squares converges, i.e., $\sum_\mathtt{t} \alpha^2_\mathtt{t}(\mathtt{s},\mathtt{a}) I(\mathtt{s}_\mathtt{t} = \mathtt{s}, \mathtt{a}_\mathtt{t} = \mathtt{a}) = O(1)$. This requirement is that the step size are not too large. It will guarantee that once we are close to the correct value, the step size will be small enough that we actually converge, and not bounce around.

Consider of the following experiment. Suppose from some time $\tau$ we update using only $Q^*$, then we clearly would like to converge. For simplicity assume there is no noise, i.e., $\omega_\mathtt{t} = 0$ for $\mathtt{t} \ge \tau$ and assume a single state., i.e., $Q^* \in \mathbb{R}$. For the update we have that $Q_{\mathtt{t}+1} = (1-\alpha_\mathtt{t})Q_\mathtt{t} + \alpha_\mathtt{t} Q^*$ or equivalently, $Q_{\mathtt{t}+1} = \beta Q_\tau + (1-\beta)Q^*$, where $\beta = \prod_{i=\tau}^{\mathtt{t}} (1-\alpha_i)$. We like to have that $\beta$ converges to 0 and a sufficient condition is that $\sum_i \alpha_i = \infty$. The small enough step size will guarantee that we will converge even when the noise is not zero (but bounded).

Many times step size are simply a function of the number of visits to $(\mathtt{s},\mathtt{a})$, which we denote by $n(\mathtt{s},\mathtt{a})$, and this widely used in practice. Two leading examples are:

1. *Linear step size:* $\alpha_{\texttt{t}}(\texttt{s},\texttt{a}) = 1/n(\texttt{s},\texttt{a})$. We have that $\sum_{n=1}^{N} 1/n = \ln(N)$ and therefore $\sum_{n=1}^{\infty} 1/n = \infty$. Also, $\sum_{n=1}^{\infty} 1/n^2 = \pi^2/6 = O(1)$

2. *Polynomial step size*: For $\theta \in (1/2, 1)$ we have $\alpha_{\texttt{t}}(\texttt{s},\texttt{a}) = 1/(n(\texttt{s},\texttt{a}))^{\theta}$. We have that $\sum_{n=1}^{N} 1/n^{\theta} \approx (1-\theta)^{-1} N^{1-\theta}$ and therefore $\sum_{n=1}^{\infty} 1/n^{\theta} = \infty$. Also, $\sum_{n=1}^{\infty} 1/n^{2\theta} \leq \frac{1}{2\theta-1}$, since $2\theta > 1$.

The linear step size, although many times popular in practice, might lead to slow converges. Here is a simple example. We have a single state $\texttt{s}$ and single action $\texttt{a}$ and $\texttt{r}(\texttt{s},\texttt{a}) = 0$. However, suppose we start with $Q_0(\texttt{s},\texttt{a}) = 1$. We will analyze the convergence with the linear step size. Our update is,

$$Q_{\texttt{t}} = (1 - \frac{1}{\texttt{t}})Q_{\texttt{t}-1} + \frac{1}{\texttt{t}}[0 + \gamma Q_{\texttt{t}-1}] = (1 - \frac{1-\gamma}{\texttt{t}})Q_{\texttt{t}-1}$$

When we solve the recursion we get that $Q_{\texttt{t}} = \Theta(1/\texttt{t}^{1-\gamma})$. [1] This implies that for $t \leq (1/\varepsilon)^{1/(1-\gamma)}$ we have $Q_{\texttt{t}} \geq \varepsilon$.

In contrast, if we use a polynomial step size, we have,

$$Q_{\texttt{t}} = (1 - \frac{1}{t^{\theta}})Q_{\texttt{t}-1} + \frac{1}{t^{\theta}}[0 + \gamma Q_{\texttt{t}-1}] = (1 - \frac{1-\gamma}{t^{\theta}})Q_{\texttt{t}-1}$$

When we solve the recursion we get that $Q_{\texttt{t}} = \Theta(e^{-(1-\gamma)t^{1-\theta}})$. This implies that for $t \geq \frac{1}{1-\gamma} \log^{1/(1-\theta)}(1/\varepsilon)$ we have $Q_{\texttt{t}} \leq \varepsilon$. This is a poly-logarithmic dependency on $\varepsilon$, which is much better. Also, not that $\theta$ is under our control, and we can set for example $\theta = 2/3$. Note that unlike $\theta$, the setting of the discount factor $\gamma$ has a huge influence on the objective function and the effective horizon.

### 9.2.6 SARSA: on-policy $Q$-learning

The Q-learning algorithm that we presented is an off-policy algorithm. This implies that it has no control over the action selection. The benefit is that it does not face the exploration exploitation trade-off and its only goal is to approximate the optimal $Q$ function.

In this section we would like to extend the $Q$-learning algorithm to an on-policy setting. This will first of all involve selecting the actions by the algorithm. Given that the actions are set by the algorithm, we can consider the return of the algorithm. We would like the return of the algorithm to converge to the optimal return.

[1] $Q_{\texttt{t}} = \prod_{i=1}^{\texttt{t}}(1-(1-\gamma)/i) \approx \prod_{i=1}^{\texttt{t}} e^{-(1-\gamma)/i} = e^{-\sum_{i=1}^{\texttt{t}}(1-\gamma)/i} \approx e^{-(1-\gamma)\ln t} = t^{-(1-\gamma)}$.

The specific algorithm that we present is called SARSA. The name comes from the fact that the feedback we observe $(\mathtt{s_t}, \mathtt{a_t}, \mathtt{r_t}, \mathtt{s_{t+1}}, \mathtt{a_{t+1}})$, ignoring the subscripts we have SARSA. Note that since it is an on-policy algorithm, the actions are actually under the control of the algorithm, and we would need to specify how to select them.

When designing the algorithm we need to think of two contradicting objectives in selecting the actions. The first is the need to explore, perform each action infinitely often. This implies that we need, for each state $\mathtt{s}$, to have that $E[\sum_{\mathtt{t}} \pi_{\mathtt{t}}(\mathtt{a}|\mathtt{s})] = \infty$. Then by Borel-Cantelli lemma we have with probability 1 an infinite number of times that we select action $\mathtt{a}$ in state $\mathtt{s}$ (actually, we need independence of the events, or at least a Martingale property, which holds in our case). On the other hand we would like not only our estimates to converge, as done in $Q$-learning, but also the return to be near optimal. For this we need the action selection to converge to being greedy with respect to the $Q$ function.

**The SARSA algorithm**

- *Initialize:* Set $Q_0(\mathtt{s}, \mathtt{a}) = 0$, for all $\mathtt{s}$, $\mathtt{a}$.
- At time $\mathtt{t} = 0, 1, \ldots$:
  – Observe $(\mathtt{s_t}, \mathtt{a_t}, \mathtt{r_t}, \mathtt{s_{t+1}})$.
  –Select $\mathtt{a_{t+1}} = \pi(\mathtt{s_{t+1}}; Q_{\mathtt{t}})$.
  – Update

$$Q_{\mathtt{t+1}}(\mathtt{s_t}, \mathtt{a_t}) := Q_{\mathtt{t}}(\mathtt{s_t}, \mathtt{a_t}) + \alpha_{\mathtt{t}}(\mathtt{s_t}, \mathtt{a_t})[\mathtt{r_t} + \gamma Q_{\mathtt{t}}(\mathtt{s_{t+1}}, a_{\mathtt{t+1}}) - Q_{\mathtt{t}}(\mathtt{s_t}, \mathtt{a_t})]$$

Note that SARSA selects an action $\mathtt{a_{t+1}}$ for state $\mathtt{s_{t+1}}$, dependent on the values of $Q_{\mathtt{t}}$ and not $Q_{\mathtt{t+1}}$. The update of $_{\mathtt{t}}$ is done only after we selected action $\mathtt{a_{t+1}}$ (which is performed at time $\mathtt{t} + 1$).

**Selecting the action:** As we discussed before, one of the main tasks of an on-policy algorithm is to select the actions. It would be natural to select the action is state $\mathtt{s_t}$ as a function of our current approximation $Q_{\mathtt{t}}$ of the optimal $Q$ function.

Given a state $\mathtt{s}$ and a $Q$ function $Q$, we first define the greedy action in state $\mathtt{s}$ according to $Q$ as

$$\bar{\mathtt{a}} = \arg\max_{\mathtt{a}} Q(\mathtt{s}, \mathtt{a})$$

The first idea might be to simply select the greedy action $\bar{\mathtt{a}}$, however this might be devastating. The main issue is that we are avoiding exploration. Some actions might look better due to errors, and we will continue to execute them and not gain any information about alternative actions.

For a concrete example, assume we initialize $Q_0$ to be 0. Consider an MDP with a single state and two actions $\mathtt{a}_1$ and $\mathtt{a}_2$. The reward of action $\mathtt{a}_1$ and $\mathtt{a}_2$ are a Bernoulli random variables with parameters 1/3 and 3/4, respectively. If we execute action $\mathtt{a}_1$ first and get a reward of 1, then we have $Q_1(\mathtt{s}, \mathtt{a}_1) > 0$ and $Q_1(\mathtt{s}, \mathtt{a}_2) = 0$. If we select the greedy action, we will *always* select action $\mathtt{a}_1$. We will both be sub-optimal in the return and never explore $\mathtt{a}_2$ which will result that we will not converge to $Q^*$. For this reason we would not select deterministically the greedy action.

In the following we will present two simple ways to select the action by $\pi(\mathtt{s}; Q)$ stochastically. Both ways will give all actions a non-zero probability, and thus guarantee exploration.

The *$\varepsilon_n$-greedy*, has as a parameter a sequence of $\varepsilon_n$ and selects the actions as follows. Let $n_{\mathtt{t}}(\mathtt{s})$ be the number of times state $\mathtt{s}$ was visited up to time $\mathtt{t}$. At time $\mathtt{t}$ in state $\mathtt{s}$ policy $\varepsilon_n$-greedy (1) with probability $1 - \varepsilon_n$ sets $\pi(\mathtt{s}; Q) = \bar{\mathtt{a}}$ where $n = n_{\mathtt{t}}(\mathtt{s})$, and (2) with probability $\varepsilon_n/|\mathcal{A}|$, selects $\pi(\mathtt{s}; Q) = \mathtt{a}$, for each $\mathtt{a} \in \mathcal{A}$. Common values for $\varepsilon_n$ are linear, $\varepsilon_{\mathtt{t}} = 1/n$, or polynomial, $\varepsilon_{\mathtt{t}} = 1/n^\theta$ for $\theta \in (0.5, 1)$.

The *soft-max*, has as a parameter a sequence of $\beta_{\mathtt{t}} \geq 0$ and select $\pi(\mathtt{s}; Q) = \mathtt{a}$, for each $\mathtt{a} \in \mathcal{A}$, with probability $\frac{e^{\beta_{\mathtt{t}} Q(\mathtt{s},\mathtt{a})}}{\sum_{\mathtt{a}' \in \mathcal{A}} e^{\beta_{\mathtt{t}} Q(\mathtt{s},\mathtt{a}')}}$. Note that for $\beta_{\mathtt{t}} = 0$ we get the uniform distribution and for $\beta_{\mathtt{t}} \to \infty$ we get the maximum. We would like the schedule of the $\beta_{\mathtt{t}}$ to go to infinity (become greedy) but need it to be slow enough (so that each action appears infinitely often).

**SARSA convergence:** Convergence of the $Q_{\mathtt{t}}$ to $Q^*$ would follow immediately from the basic properties of the $Q$ learning, we just need to guarantee that we explore each action infinitely often. (This would follow since any $\varepsilon$-greedy policy is aperiodic and irreducible, the steady state of the resulting Markov has a non-zero probability at each state, so each state would be visited infinitely often.) Then, the result will follow immediately from the convergence of $Q$-learning.

The convergence of the return of SARSA to that of $Q^*$ is more delicate. Recall that we do not have such a claim about $Q$-learning, since it is an off-policy method. For the convergence of the return we need to make our policy 'greedy enough', in the sense that it has enough exploration, but guarantees a high return through the greedy actions. The following lemma shows that if we have a strategy which is greedy with respect to a near optimal $Q$ function, then the policy is near optimal.

**Lemma 9.5.** *Let $Q$ such that $\|Q - Q^*\|_\infty \leq \Delta$, and $\pi$ the greedy policy w.r.t. $Q$, i.e., $\pi(\mathtt{s}) \in \arg\max_{\mathtt{a}} Q(\mathtt{s}, \mathtt{a})$. Then, $\mathcal{V}^*(\mathtt{s}) - \mathcal{V}^\pi(\mathtt{s}) \leq \frac{2\Delta}{1-\gamma}$.*

*Proof.* First, we show that for any state $\mathtt{s}$ we have $\mathcal{V}^*(\mathtt{s}) - Q^*(\mathtt{s}, \pi(\mathtt{s})) \leq 2\Delta$. Since $\|Q - Q^*\|_\infty \leq \Delta$ we have $|Q^*(\mathtt{s}, \pi(\mathtt{s})) - Q(\mathtt{s}, \pi(\mathtt{s}))| \leq \Delta$ and $|Q^*(\mathtt{s}, \mathtt{a}^*) - Q(\mathtt{s}, \mathtt{a}^*)| \leq \Delta$, where $\mathtt{a}^*$ is the optimal action in $\mathtt{s}$. This implies that $Q^*(\mathtt{s}, \mathtt{a}^*) - Q^*(\mathtt{s}, \pi(\mathtt{s})) \leq Q(\mathtt{s}, \mathtt{a}^*) - Q(\mathtt{s}, \pi(\mathtt{s})) + 2\Delta$. Since $\pi$ is greedy w.r.t. $Q$ we have $Q(\mathtt{s}, \pi(\mathtt{s})) \geq Q(\mathtt{s}, \mathtt{a}^*)$, and hence $\mathcal{V}^*(\mathtt{s}) - Q^*(\mathtt{s}, \pi(\mathtt{s})) = Q^*(\mathtt{s}, \mathtt{a}^*) - Q^*(\mathtt{s}, \pi(\mathtt{s})) \leq 2\Delta$.

Next,

$$\mathcal{V}^*(\mathtt{s}) = Q^*(\mathtt{s}, \mathtt{a}^*) \leq Q^*(\mathtt{s}, \pi(\mathtt{s})) + 2\Delta = E[\mathtt{r}_0] + \gamma E[\mathcal{V}^*(\mathtt{s}_1)] + 2\Delta,$$

where $\mathtt{r}_0 = E[R(\mathtt{s}, \pi(\mathtt{s}))]$ and $\mathtt{s}_1$ is the state reached when doing action $\pi(\mathtt{s})$ in state $\mathtt{s}$. As we role out to time $\mathtt{t}$ we have,

$$\mathcal{V}^*(\mathtt{s}) \leq E[\sum_{i=0}^{\mathtt{t}-1} \gamma^i \mathtt{r}_i] + \gamma^{\mathtt{t}} E[\mathcal{V}^*(\mathtt{s}_{\mathtt{t}})] + \sum_{i=1}^{\mathtt{t}} 2\Delta\gamma^i$$

where $\mathtt{r}_i$ is the reward in time $i$ in state $\mathtt{s}_i$, $\mathtt{s}_{i+1}$ is the state reached when doing action $\pi(\mathtt{s}_i)$ in state $\mathtt{s}_i$, and we start with $\mathtt{s}_0 = \mathtt{s}$. This implies that in the limit we have

$$\mathcal{V}^*(\mathtt{s}) \leq \mathcal{V}^\pi(\mathtt{s}) + \frac{2\Delta}{1-\gamma}$$

since $\mathcal{V}^\pi(\mathtt{s}) = E[\sum_{i=0}^{\infty} \gamma^i \mathtt{r}_i]$. □

The above lemma uses the greedy policy, but as we discussed before, we would like to add exploration. We would like to claim that if $\varepsilon$ is mall, then the difference in return between the greedy policy and the $\varepsilon$-greedy policy would be small. We will show a more general result, showing that for any policy, if we add a perturbation of $\varepsilon$ to the action selection, then the effect on the expected return is at most $O(\varepsilon \mathtt{V}_{\max})$.

Fix a policy $\pi$ and let $\pi_\varepsilon$ be a policy such that for any state $\mathtt{s}$ we have that $\|\pi(\cdot|\mathtt{s}) - \pi_\varepsilon(\cdot|\mathtt{s})\|_1 \leq \varepsilon$. Namely, there is a policy $\rho(\mathtt{a}|\mathtt{s})$ such that $\pi_\varepsilon(\mathtt{a}|\mathtt{s}) = (1-\varepsilon)\pi(\mathtt{a}|\mathtt{s}) + \varepsilon\rho(\mathtt{a}|\mathtt{s})$. Hence, at any state, with probability at least $1-\varepsilon$ policy $\pi_\varepsilon$ and policy $\pi$ use the same action selection.

**Lemma 9.6.** *Fix $\pi_\varepsilon$ and policy $\pi$ such that for any state $\mathtt{s}$ we have that $\|\pi(\cdot|\mathtt{s}) - \pi_\varepsilon(\cdot|\mathtt{s})\|_1 \leq \varepsilon$. Then, for any state $\mathtt{s}$ we have*

$$|\mathcal{V}^{\pi_\varepsilon}(\mathtt{s}) - \mathcal{V}^\pi(\mathtt{s})| \leq \frac{\varepsilon\gamma}{(1-\gamma)(1-\gamma(1-\varepsilon))}.$$

*For $\varepsilon \leq 1/2$ we have $|\mathcal{V}^{\pi_\varepsilon}(\mathtt{s}) - \mathcal{V}^\pi(\mathtt{s})| \leq 2\varepsilon \mathtt{V}_{\max}$*

*Proof.* Let $\mathtt{r}_\mathtt{t}$ be the reward of policy $\pi$ at time $\mathtt{t}$. By definition $\mathcal{V}^\pi(\mathtt{s}) = E[\sum_{\mathtt{t}=0}^\infty \gamma^\mathtt{t} \mathtt{r}_\mathtt{t}]$. The probability that policy $\pi_\varepsilon$ never deviated from $\pi$ until time $\mathtt{t}$ is $(1-\varepsilon)^\mathtt{t}$. Therefore we can lower bound the expected reward of policy $\pi_\varepsilon$ by $\mathcal{V}^{\pi_\varepsilon}(\mathtt{s}) \geq E[\sum_{\mathtt{t}=0}^\infty (1-\varepsilon)^\mathtt{t} \gamma^\mathtt{t} \mathtt{r}_\mathtt{t}]$.

Consider the difference between the expected returns,

$$\mathcal{V}^\pi(\mathtt{s}) - \mathcal{V}^{\pi_\varepsilon}(\mathtt{s}) \leq E[\sum_{\mathtt{t}=0}^\infty \gamma^\mathtt{t} \mathtt{r}_\mathtt{t}] - E[\sum_{\mathtt{t}=0}^\infty (1-\varepsilon)^\mathtt{t} \gamma^\mathtt{t} \mathtt{r}_\mathtt{t}]$$

Since the rewards are bounded, namely, $\mathtt{r}_\mathtt{t} \in [0,1]$, the difference is maximized if we set all the rewards to 1, and have

$$\mathcal{V}^\pi(\mathtt{s}) - \mathcal{V}^{\pi_\varepsilon}(\mathtt{s}) \leq \sum_{\mathtt{t}=0}^\infty \gamma^\mathtt{t} - \sum_{\mathtt{t}=0}^\infty (1-\varepsilon)^\mathtt{t} \gamma^\mathtt{t} = \frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\varepsilon)} = \frac{\varepsilon\gamma}{(1-\gamma)(1-\gamma(1-\varepsilon))}$$

Since $\mathtt{V}_{\max} = \frac{1}{1-\gamma}$ and $\gamma < 1$, for $\varepsilon \leq 1/2$ we have

$$\mathcal{V}^\pi(\mathtt{s}) - \mathcal{V}^{\pi_\varepsilon}(\mathtt{s}) \leq 2\varepsilon \mathtt{V}_{\max}$$

□

We can now combine the results and claim that SARSA with $\varepsilon$-greedy converges to the optimal policy. We will need that $\varepsilon_n$-greedy uses a sequence of $\varepsilon_n$ such that $\varepsilon_n$ converges to zero as $n$ increases. Call such a policy *monotone $\varepsilon_n$-greedy policy.*

**Theorem 9.7.** *For any $\lambda > 0$ there is a time $\tau$ such that at any time $\mathtt{t} > \tau$ the algorithm SARSA, using a monotone $\varepsilon_n$-greedy policy, plays a $\lambda$-optimal policy.*

*Proof.* Consider the sequence $\varepsilon_n$. Since it converges to zero, there exists a value $N$ such that for any $n \geq N$ we have $\varepsilon_n \leq 0.25\lambda/\mathtt{V}_{\max}$.

Since we are guaranteed that each state action is sampled infinitely often, there is a time $\tau_1$ such that each state is sampled at least $N$ times.

Since $Q_\mathtt{t}$ converges to $Q^*$, there is a time $\tau_2$ such that $\|Q_\mathtt{t} - Q^*\|_\infty \leq \Delta = 0.25\lambda(1-\gamma)$.

Set $\tau = \max\{\tau_1, \tau_2\}$. By Lemma 9.6 the difference between the $\varepsilon_n$-greedy policy and the greedy policy differs by at most $2\varepsilon_n \mathtt{V}_{\max} \leq \lambda$. By Lemma 9.5 the difference between the optimal and greedy policy is bounded by $2\Delta/(1-\gamma) = \lambda/2$. This implies that the policies played at time $\mathtt{t} > \tau$ are $\lambda$-optimal. □

**Expected SARSA algorithm:** In SARSA there are two sources of noise. One is from the environment, through the selection of the next state and the rewards, both not under our control. The other is from our policy, which selects a stochastic action. We can attempt to reduce this second part of the noise in the SARSA algorithm. The only difference will will be in the update, where we will not use the actual action selected by the SARSA policy, but will average over all potential actions.

The Expected SARSA algorithm:

- *Initialize:* Set $Q_0(\mathtt{s}, \mathtt{a}) = 0$, for all $\mathtt{s}$, $\mathtt{a}$.
- At time $\mathtt{t} = 0, 1, \ldots$ :

  – Observe $(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}, \mathtt{r}_\mathtt{t}, \mathtt{s}_{\mathtt{t}+1})$.

  –Select $\mathtt{a}_{\mathtt{t}+1} = \pi(\mathtt{s}_{\mathtt{t}+1}; Q_\mathtt{t})$.

  – Update

$$Q_{\mathtt{t}+1}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) := Q_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}) + \alpha_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})[\mathtt{r}_\mathtt{t} + \gamma E_{\mathtt{a}\sim\pi(\mathtt{s}_{\mathtt{t}+1}; Q_\mathtt{t})} Q_\mathtt{t}(\mathtt{s}_{\mathtt{t}+1}, \mathtt{a}) - Q_\mathtt{t}(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t})]$$

The main idea is that in the update of $Q_{\mathtt{t}+1}$ we replace the sampled action $\mathtt{a}_{\mathtt{t}+1}$, by an averaging over all potential $\mathtt{a}_{\mathtt{t}+1}$ using their probabilities. We can do it, since we define the policy $\pi(\mathtt{s}; Q)$ and know the probabilities. Due to the averaging, Expected SARSA reduces the variance of the estimates and hopefully converges faster.

## 9.3 Monte-Carlo Algorithm

Monte-Carlo methods learn directly from experience in a model free way. The idea is very simple. In order to estimate the value of a state under a given policy, i.e., $\mathcal{V}^\pi(\mathtt{s})$, we consider trajectories of the policy $\pi$ from state $\mathtt{s}$ and average them. The method does not assume any dependency between the different states, and does not even assume a Markovian environment, which is both a plus (less assumptions) and a minus (longer time to learn). (A non-Markovian environment could, for example, have the reward in a state depend on the number of visits to the state in the episode.) We will concentrate on the case of an episodic MDP, namely, generating finite length episodes in each trajectory. A special case of an episodic MDP is a finite horizon return, where all the episodes have the same length.

Assume we have a fixed policy $\pi$, which for each state $\mathtt{s}$ selects action $\mathtt{a}$ with probability $\pi(\mathtt{a}|\mathtt{s})$. Using $\pi$ we generate an episode $(\mathtt{s}_1, \mathtt{a}_1, \mathtt{r}_1, \ldots, \mathtt{s}_k, \mathtt{a}_k, \mathtt{r}_k)$. The observed return of the episode is $G = \sum_{i=1}^k \mathtt{r}_i$. We are interested in the expected return of an episode conditioned on the initial state, i.e., $\mathcal{V}^\pi(\mathtt{s}) = E[\sum_{i=1}^k \mathtt{r}_i | \mathtt{s}_1 = \mathtt{s}]$. Note that $k$ is a random variable, which is the length of the episode.

Fix a state $\mathtt{s}$, and assume you observed returns $G_1^{\mathtt{s}}, \ldots, G_m^{\mathtt{s}}$, all starting at state $\mathtt{s}$. The Monte-Carlo estimate for the state $\mathtt{s}$ would be $\widehat{V}^\pi(\mathtt{s}) = \frac{1}{m}\sum_{i=1}^m G_i$. The main issue that remains is how do we generate the samples $G_i^{\mathtt{s}}$ for a state $\mathtt{s}$. Clearly, if we assume we can reset the MDP to any state, we are done. However, such a *reset assumption* is not realistic in many applications. For this reason, we do not want to assume that we can reset the MDP to any state $\mathtt{s}$ and start an episode from it.

### 9.3.1 Generating the samples

**Initial state only** We use only the initial state of the episode. Namely, given an episode $(\mathtt{s}_1, \mathtt{a}_1, \mathtt{r}_1, \ldots, \mathtt{s}_k, \mathtt{a}_k, \mathtt{r}_k)$ we update only $\widehat{\mathcal{V}}^\pi(\mathtt{s}_1)$. This is clearly an unbiased estimate, but has many drawbacks. First, most likely it is not the case that every state can be an initial state, what do we do with such states. Second, it seems very wasteful, updating only a single state per episode.

**First visit** We update every state that appears in the episode, but update it only once. Given an episode $(\mathtt{s}_1, \mathtt{a}_1, \mathtt{r}_1, \ldots, \mathtt{s}_k, \mathtt{a}_k, \mathtt{r}_k)$ for each state $\mathtt{s}$ that appear in the episode, we consider the first appearance of $\mathtt{s}$, say $\mathtt{s}_j$, and update $\widehat{\mathcal{V}}^\pi(\mathtt{s})$ using $G^{\mathtt{s}} = \sum_{i=j}^k \mathtt{r}_i$. Namely, we compute the actual return from the first visit to state $\mathtt{s}$, and use it to update our approximation. This is clearly an unbiased estimator of the return from state $\mathtt{s}$, e.g. $E[G^{\mathtt{s}}] = \mathcal{V}^\pi(\mathtt{s})$.

**Every visit** We do an update during each step of in the episode. Namely, give an episode $(\mathtt{s}_1, \mathtt{a}_1, \mathtt{r}_1, \ldots, \mathtt{s}_k, \mathtt{a}_k, \mathtt{r}_k)$ for each state $\mathtt{s}_j$ that appears in the episode, we update each $\widehat{\mathcal{V}}^\pi(\mathtt{s}_j)$ using $G^{\mathtt{s}_j} = \sum_{i=j}^k \mathtt{r}_i$. Namely, we compute the actual return from every state $\mathtt{s}_j$ to the end, and use it to update our approximation. Note that a state can be updated multiple times in a single episode using this approach. We will later show that this estimator is biased, due to the dependency between different updates of the same state in the same episode.

**First versus Every visit:** To better understand the difference between first visit and every visit we consider the following simple test case. We have a two state MDP, actually a Markov Chain. In the initial state $\mathtt{s}_1$ we have a reward of 1 and with probability $1-p$ we stay in that state and with probability $p$ move to the terminating state $\mathtt{s}_2$. See Figure 9.1.

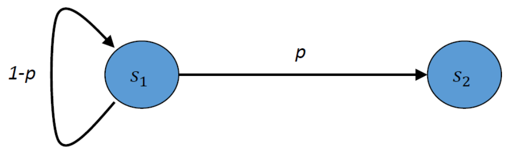


Figure 9.1: The situated agent

The expected value is $\mathcal{V}(\mathtt{s}_1) = 1/p$, which is the expected length of an episode. (Note that the return of an episode is its length, since all the rewards are 1.) Assume we observe a single trajectory, $(\mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_2)$, and all the rewards are 1. What would be a reasonable estimate for the expected return from $\mathtt{s}_1$.

`First visit` take the naive approach, considers the return from the first occurrence of $\mathtt{s}_1$, which is 4, and uses this as an estimate. `Every visit` considers four runs from state $\mathtt{s}_1$, we have: $(\mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_2)$ with return 4, $(\mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_2)$ with return 3, $(\mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_2)$ with return 2, and $(\mathtt{s}_1, \mathtt{s}_2)$ with return 1. `Every visit` averages the four and has $G = (4 + 3 + 2 + 1)/4 = 2.5$. On the face of it, the estimate of 4 seems to make more sense. We will return to this example later.

### 9.3.2 First visit

Consider the First Visit Monte-Carlo updates. Assume that for state $\mathtt{s}$ we have updates $G_1^{\mathtt{s}}, \ldots, G_m^{\mathtt{s}}$. Our estimate would be $\widehat{\mathcal{V}}^\pi(\mathtt{s}) = (1/m)\sum_{i=1}^m G_i^{\mathtt{s}}$. Since the different $G_i^{\mathtt{s}}$ are independent, we can use a concentration bound, to claim that the error is small. Actually we will need two different bounds. The first would say that if we run $n$ episodes, then with high probability we have at least $m$ episode in which state $\mathtt{s}$ appear. The second would say that if we have $m$ episodes in which state $\mathtt{s}$ appear, then we will well approximate the value function at $\mathtt{s}$. For the first part, we clearly would need to depend on the probability of reaching state $\mathtt{s}$ in an episode. Call a state $\mathtt{s}$ as $\alpha$-good if the probability that $\pi$ visits $\mathtt{s}$ in an episode is at least $\alpha$. The following theorem relates the number of episodes to the accuracy is estimating the value function.

**Theorem 9.8.** *Assume that we execute $n$ episodes using policy $\pi$ and each episode has length at most $H$. Then, with probability $1 - \delta$, for any $\alpha$-good state $\mathtt{s}$, we have $|\widehat{\mathcal{V}}^\pi(\mathtt{s}) - \mathcal{V}^\pi(\mathtt{s})| \le \lambda$, assuming $n \ge (2m/\alpha)\log(2|\mathcal{S}|/\delta)$ and $m = (H^2/\lambda^2)\log(2|\mathcal{S}|/\delta)$.*

*Proof.* Let $p(\mathtt{s})$ be the probability that policy $\pi$ visits state $\mathtt{s}$ in an episode. Since $\mathtt{s}$ is $\alpha$-good, the expected number of episodes in which $\mathtt{s}$ appears is $p(\mathtt{s})n \ge 2m\log(2|\mathcal{S}|/\delta)$.

Using the relative Chernoff–Hoffding bound (Lemma 8.2) we have that the probability that we have at least $m$ samples of state $\mathtt{s}$ is at least $1-\delta/(2|\mathcal{S}|)$.

Given that we have at least $m$ samples from state $\mathtt{s}$ using the additive Chernoff–Hoffding bound (Lemma 8.2) we have that with probability at least $1-\delta/(2|\mathcal{S}|)$ that $|\widehat{\mathcal{V}}^\pi(\mathtt{s}) - \mathcal{V}^\pi(\mathtt{s})| \leq \lambda$. (Since episodes have return in the range $[0, H]$ we need to normalize by dividing the rewards by $H$, which creates the $H^2$ term in $m$. A more refine bound can be derived by noticing that the variance of the return of an episode is bounded by $H$ and not $H^2$, and using an appropriate concentration bound, say Bernstein inequality.)

Finally, the theorem follows from a union bound over the bad events. $\square$

Next, we relate the First Visit Monte-Carlo updates to the maximum likelihood model for the MDP. Going back to the example of Figure 9.1 and observing the sequence $(\mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_2)$. The only unknown parameter is $p$.

The maximum likelihood approach would select the value of $p$ that would maximize the probability of observing the sequence $(\mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_1, \mathtt{s}_2)$. The likelihood of the sequence is, $(1-p)^3 p$. We like to solve for

$$p^* = \arg\max (1-p)^3 p$$

Taking the derivative we have $(1-p)^3 - 3(1-p)^2 p = 0$, which give $p^* = 1/4$. For the maximum likelihood (ML) model $M$ we have $p^* = 1/4$ and therefore $V(\mathtt{s}_1; M) = 4$.

In general the Maximum Likelihood model value does not always coincide with the `First Visit` Monte-carlo estimate. However we can make the following interesting connection.

Clearly, when updating state $\mathtt{s}$ using `First Visit`, we ignore all the episodes that do not include $\mathtt{s}$, and also for each the remaining episodes, that do include $\mathtt{s}$, we ignore the prefix until the fist appearance of $\mathtt{s}$. Let us modify the sample by deleting those parts (episodes in which $\mathtt{s}$ does not appear, and for each episode that $\mathtt{s}$ appears, start it at the first appearance of $\mathtt{s}$). Call this the *reduced sample*.

The maximum likelihood model, given a set of episodes, is simply the observed model. (We will not show here that the observed model is indeed the maximum likelihood model.) Namely, for each state-action pair $(\mathtt{s}, \mathtt{a})$ let $n(\mathtt{s}, \mathtt{a})$ be the number of times it appears, let $n(\mathtt{s}, \mathtt{a}, \mathtt{s}')$ be the number of times $\mathtt{s}'$ is observed following executing action $\mathtt{a}$ in state $\mathtt{s}$. The observed transition model is $\widehat{p}(\mathtt{s}'|\mathtt{s}, \mathtt{a}) = n(\mathtt{s}, \mathtt{a}, \mathtt{s}')/n(\mathtt{s}, \mathtt{a})$. Assume that in the $i$-th execution of action $\mathtt{a}$ in state $\mathtt{s}$ we observe a reward $\mathtt{r}_i$ then the observed reward is $\widehat{\mathtt{r}}(\mathtt{s}, \mathtt{a}) = (1/n(\mathtt{s}, \mathtt{a})) \sum_{i=1}^{n(\mathtt{s},\mathtt{a})} \mathtt{r}_i$.

**Theorem 9.9.** *Let $M$ be the maximum likelihood MDP for the reduced sample. The expected value of* $\mathtt{s}_0$ *in $M$, i.e., $\mathcal{V}(\mathtt{s}; M)$, is identical to the* `First Visit` *estimate of* $\mathtt{s}_0$*, i.e.,* $\widehat{\mathcal{V}}^\pi(\mathtt{s}_0)$.

*Proof.* Assume that we have $N$ episodes in the reduced sample and the sum of the rewards in the $i$-th episode is $G_i$. The First Visit Monte Carlo estimate would be $\widehat{\mathcal{V}}^\pi(\mathtt{s}_0) = (1/N)\sum_{i=1}^N G_i$.

Consider the maximum likelihood model. Since we have a fixed deterministic policy, we can ignore actions, and define $n(\mathtt{s}) = n(\mathtt{s}, \pi(\mathtt{s}))$ and $\widehat{\mathtt{r}}(\mathtt{s}, \pi(\mathtt{s})) = \widehat{\mathtt{r}}(\mathtt{s})$. We set the initial state $\mathtt{s}_0$ to be the state we are updating.

We want to compute the expected number of visits $\mu(\mathtt{s})$ to each state $\mathtt{s}$ in the ML model $M$. We will show that $\mu(\mathtt{s}) = n(\mathtt{s})/N$. This implies that the expected reward for state $\mathtt{s}_0$ in $M$ would be

$$\mathcal{V}^\pi(\mathtt{s}_0; M) = \sum_v \mu(v)\widehat{r}(v) = \sum_v \frac{n(v)}{N}\frac{1}{n(v)}\sum_{i=1}^{n(v)} \mathtt{r}_i^v = \frac{1}{N}\sum_{j=1}^{N} G_j$$

where the last equality follow by changing the order of summation (from states to episodes).

It remains to show that $\mu(\mathtt{s}) = n(\mathtt{s})/N$. We have the following identities. For $v \neq \mathtt{s}_0$:

$$\mu(v) = \sum_u \widehat{p}(v|u)\mu(u)$$

For the initial state we have

$$\mu(\mathtt{s}_0) = 1 + \sum_u \widehat{p}(\mathtt{s}_0|u)\mu(u)$$

Note that $n(v) = \sum_u n(u, v)$ for $v \neq \mathtt{s}_0$ and $n(\mathtt{s}_0) = N + \sum_u n(u, \mathtt{s}_0)$, and recall that $\widehat{p}(v|u) = n(u, v)/n(u)$. One can verify the identities by plugging in these values. □

### 9.3.3 Every visit

The `First Visit` updates are unbiased, since the different updates are from different episodes. For each episode that update is an independent unbiased sample of the return. For `Every Visit` the situation is more complicated, since there are different updates from the same episode, and therefore they are dependent. The first issue that we have to resolve is how do we like to average for the `Every Visit` updates. Let $G_{i,j}^{\mathtt{s}}$ be the $j$-th update in the $i$-th episode for state $\mathtt{s}$. Let $n_i$ be the number of updates in episode $i$ and $N$ the overall number of episodes.

One way to average the updates is to average for each episode the updates and average across episodes. Namely,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}G_{i,j}$$

An alternative approach is to sum the updates and divide by the number of updates,

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}G_{i,j}}{\sum_{i=1}^{N}n_i}$$

We will use the latter scheme, but it is worthwhile understanding the difference between the two. Consider for example the case that we have 10 episodes, in 9 we have a single visit to `s` and a return of 1, and in the 10-th we have 11 visits to `s` and all the returns are zero. The first averaging would give an estimate of 9/10 while the second would give an estimate of 9/20.

Consider the case of Figure 9.1. For a single episode of length $k$ we have that the sum of the rewards is $k(k+1)/2$, since there are updates of lengths $k,\ldots,1$ and recall that the return equals the length since all rewards are 1. The number of updates is $k$, so we have that the estimate of a single episode is $(k+1)/2$. When we take the expectation we have that $E[(k+1)/2]=(1/p+1)/2$ which is different from the expected value of $1/p$. (Recall that the `Every Visit` updates $k$ times using values $k,\ldots,1$. In addition, $E[k]=1/p$ which is also the expected value.) If we have a single episode then both averaging schemes are identical.

When we have multiple episodes, we can see the difference between the two averaging schemes. The first will be biased random variables of $E[(k+1)/2]=(1/p+1)/2$, so it will converge to this value rather than $1/p$. The second scheme, which we will use in `Every Visit` updates, will have the bias decrease with the number of episodes. The reason is that we sum separately the returns, and the number of occurrences. This implies that we have

$$E[\mathcal{V}^{ev}(\mathtt{s}_1)]=\frac{E[k(k+1)/2]}{E[k]}=\frac{1}{p},$$

since $E[k^2]=2/p^2-1/p$. This implies that if we average many episodes we will get an almost unbiased estimate using `Every Visit`.

We did all this on the example of Figure 9.1, but this indeed generalizes. Given an arbitrary episodic MDP, consider the following mapping. For each episode, mark the places where state `s` appears (the state we want to approximate its value). We

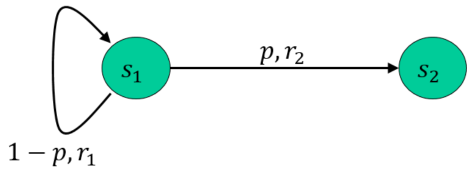


Figure 9.2: The situated agent

now have a distribution of rewards from going from $\mathtt{s}$ back to $\mathtt{s}$. Since we are in an episodic MDP, we also have to terminate, and for this we can add another state, from which we transition from state $\mathtt{s}$ and have the reward distribution as the rewards from the last appearance of $\mathtt{s}$ until the end of the episode. This implies that we have two states MDP as described in Figure 9.2.

For this MDP, the value is $\mathcal{V}^\pi(\mathtt{s}_1) = \frac{1-p}{p}\mathtt{r}_1 + \mathtt{r}_2$. The single episode expected estimate of `Every Visit` is $\mathcal{V}^\pi(\mathtt{s}_1) = \frac{1-p}{2p}\mathtt{r}_1 + \mathtt{r}_2$. The $m$ episodes expected estimate of `Every Visit` is $V(\mathtt{s}_1) = \frac{m}{m+1}\frac{1-p}{p}\mathtt{r}_1 + \mathtt{r}_2$. This implies that if we have a large number of episodes the bias of the estimate becomes negligible. (For more details, see Theorem 7 in [12].)

### Every visit and squared loss

Recall that the squared error is summing over all the observation the squared error. Assume we have $\mathtt{s}_{i,j}$ as the $j$-th state in the $i$-th episode, and it has return $G_{i,j}$, i.e., the sum of the rewards from in episode $i$ from step $j$ until the end of the episode. Let $\widehat{V}(\mathtt{s}_{i,j})$ be the estimate that `Every Visit` for state $\mathtt{s}_{i,j}$. (Note that states $\mathtt{s}_{i,j}$ are not unique, and we can have $\mathtt{s} = \mathtt{s}_{i_1,j_1} = \mathtt{s}_{i_2,j_2}$.) The square error is

$$SE = \frac{1}{2}\sum_{i,j}(\widehat{V}(\mathtt{s}_{i,j}) - G_{i,j})^2$$

For a fixed state $\mathtt{s}$ we have

$$SE(\mathtt{s}) = \frac{1}{2}\sum_{i,j:\mathtt{s}=\mathtt{s}_{i,j}}(\widehat{V}(\mathtt{s}) - \mathtt{r}_{i,j})^2$$

and the total squared error is $SE = \sum_{\mathtt{s}} SE(\mathtt{s})$.

Our goal is to select a value $\widehat{V}^{se}(\mathtt{s})$ for every state, which would minimize the SE. The minimization is achieved by minimizing the square error of each $\mathtt{s}$, and setting the values

$$\widehat{V}^{se}(\mathtt{s}) = \frac{\sum_{i,j:\mathtt{s}=\mathtt{s}_{i,j}} G_{i,j}}{|(i,j) : \mathtt{s} = \mathtt{s}_{i,j}|},$$

which is exactly the `Every Visit` Monte-carlo estimate for state $\mathtt{s}$.

### 9.3.4 Monte-Carlo control

We can also use the Monte-Carlo methodology to learn the optimal policy. The main idea is to learn the $Q^{\pi}$ function. This is done by simply updating for every $(\mathtt{s}, \mathtt{a})$. (The updates can be either `Every Visit` or `First Visit`.) The problem is that we need the policy to be "exploring", otherwise we will not have enough information about the actions the policy does not perform.

For the control, we can maintain an estimates of the $Q^{\pi}$ function, where the current policy is $\pi$. After we have a good estimate of $Q^{\pi}$ we can switch to a policy which is greedy with respect to $Q^{\pi}$. Namely, each time we reach a state $\mathtt{s}$, we select a "near-greedy" action, for example, use $\varepsilon$-greedy.

We will show that updating from one $\varepsilon$-greedy policy to another $\varepsilon$-greedy policy, using policy improvement, does increase the value of the policy. This will guarantee that we will no cycle, and eventually converge.

Recall that an $\varepsilon$-greedy policy, can be define in the following way. For every state $\mathtt{s}$ there is an action $\bar{\mathtt{a}}_{\mathtt{s}}$, which is the preferred action. The policy does the following: (1) with probability $1 - \varepsilon$ selects action $\bar{\mathtt{a}}$. (2) with probability $\varepsilon$, selects each action $\mathtt{a} \in \mathcal{A}$, with probability $\varepsilon / |\mathcal{A}|$.

Assume we have an $\varepsilon$-greedy policy $\pi_1$. Compute $Q^{\pi_1}$ and define $\pi_2$ to be $\varepsilon$-greedy with respect to $Q^{\pi_1}$.

**Theorem 9.10.** *For any $\varepsilon$-greedy policy $\pi_1$, the $\varepsilon$-greedy improvement policy $\pi_2$ has $\mathcal{V}^{\pi_2} \geq \mathcal{V}^{\pi_1}$.*

*Proof.* Let $\bar{\mathtt{a}}_{\mathtt{s}} = \arg\max_{\mathtt{a}} Q^{\pi_1}(\mathtt{s}, \mathtt{a})$ be the greedy action w.r.t. $Q^{\pi_1}$. We now lower bound the value of $Q^{\pi_2}$.

$$\begin{aligned}
E_{\mathtt{a}\sim\pi_2(\cdot|\mathtt{s})}[Q^{\pi_1}(\mathtt{s},\mathtt{a})] &= \sum_{\mathtt{a}\in\mathcal{A}} \pi_2(\mathtt{a}|\mathtt{s})Q^{\pi_1}(\mathtt{s},\mathtt{a}) \\
&= \frac{\varepsilon}{|\mathcal{A}|}\sum_{\mathtt{a}\in\mathcal{A}} Q^{\pi_1}(\mathtt{s},\mathtt{a}) + (1-\varepsilon)Q^{\pi_1}(\mathtt{s},\bar{\mathtt{a}}_{\mathtt{s}}) \\
&\geq \frac{\varepsilon}{|\mathcal{A}|}\sum_{\mathtt{a}\in\mathcal{A}} Q^{\pi_1}(\mathtt{s},\mathtt{a}) + (1-\varepsilon)\sum_{\mathtt{a}\in\mathcal{A}}\frac{\pi_1(\mathtt{a}|\mathtt{s})-\varepsilon/|\mathcal{A}|}{1-\varepsilon}Q^{\pi_1}(\mathtt{s},\mathtt{a}) \\
&= \sum_{\mathtt{a}\in\mathcal{A}} \pi_1(\mathtt{a}|\mathtt{s})Q^{\pi_1}(\mathtt{s},\bar{\mathtt{a}}) = \mathcal{V}^{\pi_1}(\mathtt{s})
\end{aligned}$$

The inequality follows, since we are essentially concentrating of the action that $\pi_1(\cdot|\mathtt{s})$ selects with probability $1-\varepsilon$, and clearly $\bar{\mathtt{a}}_{\mathtt{s}}$, by definition, guarantees a higher value.

It remains to show, similar to the basic policy improvement, that we have

$$\mathcal{V}^{\pi_2} \geq E_{\mathtt{a}\sim\pi_2(\cdot|\mathtt{s})}[Q^{\pi_1}(\mathtt{s},\mathtt{a})] \geq \mathcal{V}^{\pi_1}.$$

Basically, we need to re-write the Bellman optimality operator to apply to $\varepsilon$-greedy policies as follows:

$$(T^*_\varepsilon V)(\mathtt{s}) = \max_{\mathtt{a}}[(1-\varepsilon)(\mathtt{r}(\mathtt{s},\mathtt{a})+\gamma E_{\mathtt{s}'\sim p(\cdot|\mathtt{s},\mathtt{a})}[V(\mathtt{s}')])+\varepsilon(\sum_{\mathtt{a}''}\frac{\varepsilon}{|\mathcal{A}|}(\mathtt{r}(\mathtt{s},\mathtt{a}'')+\gamma E_{\mathtt{s}''\sim p(\cdot|s,\mathtt{a}'')}[V(\mathtt{s}'')]))]$$

Clearly $T^*_\varepsilon(V)$ is monotone in $V$, and for $\mathcal{V}^{\pi_1}$ we have $T^*_\varepsilon(\mathcal{V}^{\pi_1}) = T^{\pi_2}(\mathcal{V}^{\pi_1})$. Since $T^{\pi_2}(\mathcal{V}^{\pi_1}) = E_{\mathtt{a}\sim\pi_2(\cdot|\mathtt{s})}[Q^{\pi_1}(\mathtt{s},\mathtt{a})]$, this implies that,

$$T^{\pi_2}(\mathcal{V}^{\pi_1}) = T^*_\varepsilon(\mathcal{V}^{\pi}_1) \geq T^{\pi_1}(\mathcal{V}^{\pi_1}) = \mathcal{V}^{\pi_1}$$

We can continue to apply $T^{\pi_2}$ and due to the monotonicity have

$$(T^{\pi_2})^k(\mathcal{V}^{\pi_1}) \geq (T^{\pi_2})^{k-1}(\mathcal{V}^{\pi_1}) \geq \cdots \geq \mathcal{V}^{\pi_1}$$

Since $\lim_{k\to\infty}(T^{\pi_2})^k(\mathcal{V}^{\pi_1}) = \mathcal{V}^{\pi_2}$ we are done. □

### 9.3.5 Monte-Carlo: pros and cons

The main benefits of the Monte-Carlo updates are:

1. Very simple and intuitive

2. Does not assume the environment is Markovian
3. Extends naturally to function approximation (more in future chapters)
4. Unbiased updates (using `First Visit`).

The main drawback of the Monte-Carlo updates are:

1. Need to wait for the end of episode to update.
2. Suited mainly for episodic environment.
3. Biased updates (using `Every Visit`).

## 9.4 Temporal Difference algorithms

In this section we will look on temporal differences methods, which works in an online fashion. We will start with $TD(0)$ which uses only the most recent observations for the updates, and we will continue with methods that allow for a longer look-ahead, and then consider $TD(\lambda)$ which averages multiple look-ahead estimations.

In general, temporal differences (TD) methods, learn directly from experience, and therefore are model-free methods. Unlike Monte-Carlo algorithms, they will use incomplete episodes for the updates, and they are not restricted to episodic MDPs. The TD methods update their estimates given the current observation and in that direction, similar in spirit to $Q$-learning and SARSA.

### 9.4.1 TD(0)

Fix a policy $\pi \in SD$, stationary and deterministic. The goal is to learn the value function $\mathcal{V}^\pi(\mathtt{s})$ for every $s \in S$. (The same goal as Monte-Carlo learning.) The TD algorithms will maintain an estimate of the value function of the policy $\pi$, i.e., maintain an estimate $\widehat{V}_\mathtt{t}(\mathtt{s})$ for $\mathcal{V}^\pi(\mathtt{s})$. The TD algorithms will use their estimates $\widehat{V}$ for the updates. This implies that unlike Monte-Carlo, there will be an interaction between the estimates of different states and at different times.

As a starting point, we can recall the *value iteration* algorithm.

$$V_{\mathtt{t}+1}(\mathtt{s}) = E^\pi[\mathtt{r}(\mathtt{s}, \pi(\mathtt{s})) + \gamma V_\mathtt{t}(\mathtt{s}')]$$

We have shown that value iteration converges, namely $V_\mathtt{t} \to_{\mathtt{t}\to\infty} \mathcal{V}^\pi$.

Assume we sample $(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}, \mathtt{r}_\mathtt{t}, \mathtt{s}_{\mathtt{t}+1})$. Let $\widehat{V}_\mathtt{t}$ our estimation at time $\mathtt{t}$, and we sample $(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}, \mathtt{r}_\mathtt{t}, \mathtt{s}_{\mathtt{t}+1})$. Then,

$$E^\pi[\widehat{V}_\mathtt{t}(\mathtt{s}_\mathtt{t})] = E^\pi[\mathtt{r}_\mathtt{t} + \gamma \widehat{V}_\mathtt{t}(\mathtt{s}_{\mathtt{t}+1})] = E^\pi[\mathtt{r}(\mathtt{s}, \mathtt{a}) + \gamma \widehat{V}_\mathtt{t}(\mathtt{s}') | \mathtt{s} = \mathtt{s}_\mathtt{t}, a = \pi(\mathtt{s})]$$

The $TD(0)$ will do an update in this direction, namely, $[\mathtt{r}_\mathtt{t} + \gamma \widehat{V}_\mathtt{t}(\mathtt{s}_{\mathtt{t}+1})]$.

$TD(0)$ **Algorithm**

- *Initialize:* $\widehat{V}(\mathtt{s})$ arbitrarily.
- At time $\mathtt{t} = 0, 1, \ldots$:
  - Observe: $(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}, \mathtt{r}_\mathtt{t}, \mathtt{s}_{\mathtt{t}+1})$.

– Update:

$$\widehat{V}(\mathtt{s_t}) = \widehat{V}(\mathtt{s_t}) + \alpha_{\mathtt{t}}(\mathtt{s_t}, \mathtt{a_t})\big[\mathtt{r_t} + \gamma\widehat{V}(\mathtt{s_{t+1}}) - \widehat{V}(\mathtt{s_t})\big]$$

where $\alpha_{\mathtt{t}}(\mathtt{s}, \mathtt{a})$ is the step size for $(\mathtt{s}, \mathtt{a})$ at time $\mathtt{t}$.

We define the *temporal difference* to be

$$\Delta_{\mathtt{t}} = \mathtt{r_t} + \gamma\widehat{V}(\mathtt{s_{t+1}}) - \widehat{V}(\mathtt{s_t})$$

The $TD(0)$ update becomes:

$$\widehat{V}(\mathtt{s_t}) = \widehat{V}(\mathtt{s_t}) + \alpha_{\mathtt{t}}(\mathtt{s_t}, \mathtt{a_t})\Delta_{\mathtt{t}}$$

[[YM: motivating example. Waze?!]]

We would like to compare the $TD(0)$ and the Monte-Carlo (MC) algorithms. Here is a simple example with four states $S = \{A, B, C, D\}$ where $\{C, D\}$ are terminal states and in $\{A, B\}$ there is one action (essentially, the policy selects a unique action). Assume we observe eight episodes. One episode is $(A, 0, B, 0, C)$, one episode $(B, 0, C)$, and six episodes $(B, 1, D)$. We would like to estimate the value function of the non-terminal states. For $V(B)$ both $TD(0)$ and $MC$ will give $6/8 = 0.75$. The interesting question would be: what is the estimate for $A$? MC will average only the trajectories that include $A$ and will get 0 (only one trajectory which gives 0 reward). The $TD(0)$ will consider the value from $B$ as well, and will give an estimate of 0.75. (Assume that the $TD(0)$ continuously updates using the same episodes until it converges.)

We would like to better understand the above example. For the above example the empirical MDP will have a transition from $A$ to $B$, with probability 1 and reward 0, from $B$ we will have a transition to $C$ with probability 0.25 and reward 0 and a transition to $D$ with probability 0.75 and reward 1. (See, Figure 9.3.) The value of $A$ in the empirical model is 0.75. In this case the empirical model agrees with the $TD(0)$ estimate, we show that this holds in general.

**Maximum Likelihood model** Recall that the empirical model, which is also the maximum likelihood model, of a sample be define as follows. Let $n(\mathtt{s}, \mathtt{a})$ be the number of times $(\mathtt{s}, \mathtt{a})$ appears in the sample. Given a sample $(\mathtt{s_t}, \mathtt{a_t}, \mathtt{r_t}, \mathtt{s_{t+1}})$ for $1 \le t \le T$, we define the empirical model as follows. The rewards $\widehat{\mathtt{r}}(\mathtt{s}, \mathtt{a})$ are the average rewards of $(\mathtt{s}, \mathtt{a})$, i.e., $\widehat{\mathtt{r}}(\mathtt{s}, \mathtt{a}) = \frac{1}{n(\mathtt{s},\mathtt{a})} \sum_{\mathtt{t}:\mathtt{s_t}=\mathtt{s},\mathtt{a_t}=\mathtt{a}} \mathtt{r_t}$ and $\widehat{p}(\mathtt{s}'|\mathtt{s}, \mathtt{a}) = \frac{n(\mathtt{s},\mathtt{a},\mathtt{s}')}{n(\mathtt{s},\mathtt{a})}$, where $n(\mathtt{s}, \mathtt{a}, \mathtt{s}')$ is the number of samples that have $\mathtt{s_{t+1}} = \mathtt{s}'$ when $\mathtt{s_t} = \mathtt{s}$ and $\mathtt{a_t} = \mathtt{a}$.

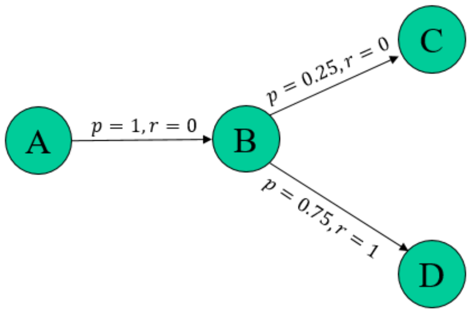


Figure 9.3: The situated agent

The following theorem states that the value of $\pi$ on the empirical model is identical to that of $TD(0)$ (running on the sample until convergence, namely, continuously sampling uniformly $t \in [1, T]$ and using $(\mathtt{s_t}, \mathtt{a}_r, \mathtt{r_t}, \mathtt{s_{t+1}})$ for the $TD(0)$ update).

**Theorem 9.11.** *Let $\mathcal{V}^{\pi}_{TD}$ be the estimated value function of $\pi$ when we run $TD(0)$ until convergence. let $\mathcal{V}^{\pi}_{EM}$ be the value function of $\pi$ on the empirical model. Then $\mathcal{V}^{\pi}_{TD} = \mathcal{V}^{\pi}_{EM}$.*

*Proof sketch.* The update of $TD(0)$ is $\widehat{V}(\mathtt{s_t}) = \widehat{V}(\mathtt{s_t}) + \alpha_{\mathtt{t}}(\mathtt{s_t}, \mathtt{a_t})\Delta_{\mathtt{t}}$, where $\Delta_{\mathtt{t}} = \mathtt{r_t} + \gamma \widehat{V}(\mathtt{s_{t+1}}) - \widehat{V}(\mathtt{s_t})$. At convergence we have $E[\Delta_{\mathtt{t}}] = 0$ and hence,

$$\widehat{V}(\mathtt{s}) = \frac{1}{n(\mathtt{s}, \mathtt{a})} \sum_{\mathtt{s_{t+1}} : \mathtt{s_t} = \mathtt{s}, \mathtt{a_t} = \mathtt{a}} \mathtt{r_t} + \gamma \widehat{V}(\mathtt{s_{t+1}}) = \widehat{\mathtt{r}}(\mathtt{s}, \mathtt{a}) + \gamma E_{\mathtt{s'} \sim \widehat{p}(\cdot|\mathtt{s}, \mathtt{a})}[\widehat{V}(\mathtt{s'})]$$

where $\mathtt{a} = \pi(\mathtt{s})$. □

In is worth to compare the above theorem to the case of Monte Carlo (Theorem 9.9). Here we are using the entire sample, and we have the same ML model for any state $\mathtt{s}$. In the Monte-Carlo case we used a reduced sample, which depends on the state $\mathtt{s}$ and therefore we have a different ML model for each state, based on its reduced sample.

**Convergence:** The proof of the convergence of $TD(0)$ is very similar to that of $Q$-learning. We will show that $TD(0)$ is an instance of the stochastic approximation algorithm, as presented in previously in Section 9.2.3, and the convergence proof will follow from this.

**Theorem 9.12** (Convergence $TD(0)$)**.** *If the step size has the properties that for every* $(\mathtt{s},\mathtt{a})$ *we have* $\sum_{\mathtt{t}} \alpha_{\mathtt{t}}(\mathtt{s},\mathtt{a}) = \infty$ *and* $\sum_{\mathtt{t}} \alpha^2_{\mathtt{t}}(\mathtt{s},\mathtt{a}) = O(1)$*, then* $\widehat{V}$ *converges to* $\mathcal{V}^\pi$*, with probability* 1.

We will show the convergence using the general theorem for stochastic approximation iterative algorithm (Theorem 9.4).

We first define a linear operator $H$ for the policy $\pi$,

$$(Hv)(\mathtt{s}) = \mathtt{r}(\mathtt{s},\pi(\mathtt{s})) + \gamma \sum_{\mathtt{s}'} p(\mathtt{s}'|\mathtt{s},\pi(\mathtt{s}))v(\mathtt{s}')$$

Note that $H$ is the operator $T^\pi$ we define in Section 5.4.3.

As we saw before, the operator $H$ is a $\gamma$-contracting operator

$$\begin{aligned}\|Hv_1 - Hv_2\|_\infty &= \gamma \max_s |\sum_{\mathtt{s}'} p(\mathtt{s}'|\mathtt{s},\pi(\mathtt{s}))(v_1(\mathtt{s}') - v_2(\mathtt{s}'))| \\ &\leq \gamma \max_{\mathtt{s}'} |v_1(\mathtt{s}') - v_2(\mathtt{s}')| \\ &\leq \gamma \|v_1 - v_2\|_\infty\end{aligned}$$

We now would like to re-write the $TD(0)$ update to be a stochastic approximation iterative algorithm. The $TD(0)$ update is,

$$V_{\mathtt{t}+1}(\mathtt{s}_{\mathtt{t}}) = (1-\alpha_{\mathtt{t}})V_{\mathtt{t}} + \alpha_{\mathtt{t}}\Phi_{\mathtt{t}}$$

where $\Phi_{\mathtt{t}} = \mathtt{r}_{\mathtt{t}} + \gamma V_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}+1})$. We would like to consider the expected value of $\Phi_{\mathtt{t}}$. Clearly, $E[\mathtt{r}_{\mathtt{t}}] = \mathtt{r}(\mathtt{s}_{\mathtt{t}},\pi(\mathtt{s}_{\mathtt{t}}))$ and $\mathtt{s}_{\mathtt{t}+1} \sim p(\cdot|\mathtt{s}_{\mathtt{t}},\mathtt{a}_{\mathtt{t}})$. This implies that $E[\Phi_{\mathtt{t}}] = (HV_{\mathtt{t}})(\mathtt{s}_{\mathtt{t}})$. Therefore, we can define the noise term $\omega_{\mathtt{t}}$ as follows,

$$\omega_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}) = [\mathtt{r}_{\mathtt{t}} + \gamma V_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}+1})] - (HV_{\mathtt{t}})(\mathtt{s}_{\mathtt{t}})\ ,$$

and have $E[\omega_{\mathtt{t}}] = 0$. We can bound $|\omega_{\mathtt{t}}| \leq V_{max} = \frac{R_{max}}{1-\gamma}$, since the value function is bounded by $V_{max}$.

Returning to $TD(0)$, we can write

$$V_{\mathtt{t}+1}(\mathtt{s}_{\mathtt{t}}) = (1-\alpha_{\mathtt{t}})V_{\mathtt{t}} + \alpha_{\mathtt{t}}[(HV_{\mathtt{t}})(\mathtt{s}_{\mathtt{t}}) + \omega_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}})]$$

The requirement of the step size appears in the statement of the theorem (and so holds). The noise $\omega_{\mathtt{t}}$ has both $E[\omega_{\mathtt{t}}] = 0$ and $|\omega_{\mathtt{t}}| \leq V_{max}$. And the operator $H$ is contracting with a fix-point $\mathcal{V}^\pi$. Therefore, using Theorem 9.4, we established Theorem 9.12.

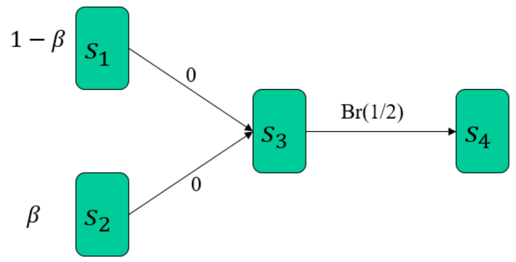


Figure 9.4: Markov Reward Chain

**Comparing various algorithms:** When we compare $TD(0)$, to $MC$ and both to Dynamic Programming (DP) we can view it as different ways of computing the value function.

- MC We have $\mathcal{V}^\pi(\mathtt{s}) = E[R_\mathtt{t}|\mathtt{s}_\mathtt{t} = \mathtt{s}]$. In MC we observe the return, $R_\mathtt{t}$, of episodes, and their mean is what we like to estimate.
- $TD(0)$ We have $\mathcal{V}^\pi(\mathtt{s}) = E[\mathtt{r}_\mathtt{t} + \gamma\mathcal{V}^\pi(\mathtt{s}_{\mathtt{t}+1})|\mathtt{s}_\mathtt{t} = \mathtt{s}]$. In $TD(0)$ we observe samples of $\mathtt{r}_\mathtt{t}$, we use our estimate for $\mathcal{V}^\pi(\mathtt{s}_{\mathtt{t}+1})$ and the expectation is what we like to estimate (assuming our estimates converge).
- DP We have $\mathcal{V}^\pi(\mathtt{s}) = \sum_\mathtt{a} \pi(\mathtt{a}|\mathtt{s})[\mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma \sum_{\mathtt{s}'} p(\mathtt{s}'|\mathtt{s},\mathtt{a})\mathcal{V}^\pi(\mathtt{s}')]$. In DP we have the entire model, and use it to compute expectations.

We can see the difference between $TD(0)$ and $MC$ in the Markov Chain in Figure 9.4. To get an approximation of state $\mathtt{s}_2$, i.e., $|\widehat{V}(\mathtt{s}_2) - \frac{1}{2}| \approx \varepsilon$. The Monte-Carlo will require $O(1/(\beta\varepsilon^2))$ episodes (out of which only $O(1/\varepsilon^2)$ start at $\mathtt{s}_2$) and the $TD(0)$ will require only $O(1/\varepsilon^2 + 1/\beta)$ since the estimate of $\mathtt{s}_3$ will converge after $1/\varepsilon^2$ episodes which start from $\mathtt{s}_1$.

### 9.4.2 TD: Multiple look-ahead

The $TD(0)$ uses only the current reward and state. Given $(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}, \mathtt{r}_\mathtt{t}, \mathtt{s}_{\mathtt{t}+1})$ it updates $\Delta_\mathtt{t} = R^{(1)}_\mathtt{t}(\mathtt{s}_\mathtt{t}) - \widehat{V}(\mathtt{s}_\mathtt{t})$ where $R^{(1)}_\mathtt{t}(\mathtt{s}_\mathtt{t}) = \mathtt{r}_\mathtt{t} + \gamma\widehat{V}(\mathtt{s}_{\mathtt{t}+1})$. We can also consider a two step look-ahead as follows. Given $(\mathtt{s}_\mathtt{t}, \mathtt{a}_\mathtt{t}, \mathtt{r}_\mathtt{t}, \mathtt{s}_{\mathtt{t}+1}, \mathtt{a}_{\mathtt{t}+1}, \mathtt{r}_{\mathtt{t}+1}, \mathtt{s}_{\mathtt{t}+2})$ we can update using $\Delta^{(2)}_\mathtt{t} = R^{(2)}_\mathtt{t}(\mathtt{s}_\mathtt{t}) - \widehat{V}(\mathtt{s}_\mathtt{t})$ where $R^{(2)}_\mathtt{t}(\mathtt{s}_\mathtt{t}) = \mathtt{r}_\mathtt{t} + \gamma\mathtt{r}_{\mathtt{t}+1} + \gamma^2\widehat{V}(\mathtt{s}_{\mathtt{t}+2})$. Using the same logic, we have that this is a temporal difference that uses a two time steps.

We can generalize this to any $n$-step look-ahead and define $R_{\mathtt{t}}^{(n)}(\mathtt{s_t}) = \sum_{i=0}^{n-1}\gamma^i \mathtt{r}_{\mathtt{t}+i} + \gamma^n \widehat{V}(\mathtt{s}_{\mathtt{t}+n})$ and updates $\Delta_{\mathtt{t}}^{(n)} = R_{\mathtt{t}}^{(n)}(\mathtt{s_t}) - \widehat{V}(\mathtt{s_t})$.

We can relate the $\Delta_{\mathtt{t}}^{(n)}$ to the $\Delta_{\mathtt{t}}$ as follows:

$$\Delta_{\mathtt{t}}^{(n)} = \sum_{i=0}^{n-1}\gamma^i \Delta_{\mathtt{t}+i}$$

This follows since

$$\sum_{i=0}^{n-1}\gamma^i \Delta_{\mathtt{t}+i} = \sum_{i=0}^{n-1}\gamma^i \mathtt{r}_{\mathtt{t}+i} + \sum_{i=0}^{n-1}\gamma^{i+1}\widehat{V}(\mathtt{s}_{\mathtt{t}+i+1}) - \sum_{i=0}^{n-1}\gamma^i \widehat{V}(\mathtt{s}_{\mathtt{t}+i}) = \sum_{i=0}^{n-1}\gamma^i \mathtt{r}_{\mathtt{t}+i} + \gamma^n \widehat{V}(\mathtt{s}_{\mathtt{t}+n}) - \widehat{V}(\mathtt{s_t})$$

Using the $n$-step look-ahead we have $\widehat{V}(\mathtt{s_t}) = \widehat{V}(\mathtt{s_t}) + \alpha_{\mathtt{t}}\Delta_{\mathtt{t}}^{(n)}$ where $\Delta_{\mathtt{t}}^{(n)} = R_{\mathtt{t}}^{(n)}(\mathtt{s_t}) - \widehat{V}(\mathtt{s_t})$. Note that the operator $R_{\mathtt{t}}^{(n)}$ is $\gamma^n$-contracting, namely

$$\|R_{\mathtt{t}}^{(n)}(V_1) - R_{\mathtt{t}}^{(n)}(V_2)\|_\infty \le \gamma^n \|V_1 - V_2\|_\infty$$

We can use any parameter $n$ for the $n$-step look-ahead. If the episode ends before step $n$ we can pad it with rewards zero. This implies that for $n = \infty$ we have that $n$-step look-ahead is simply the Monte-Carlo estimate. However, we need to select some parameter $n$. An alternative idea is to simply average over the possible parameters $n$. One simple way to average is to use exponential averaging with a parameter $\lambda \in (0,1)$. This implies that the weight of each parameter $n$ is $(1-\lambda)\lambda^{n-1}$.

This leads us to the $TD(\lambda)$ update:

$$\widehat{V}(\mathtt{s_t}) = \widehat{V}(\mathtt{s_t}) + \alpha_{\mathtt{t}}(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\Delta_{\mathtt{t}}^{(n)}$$

**Remark:** While both $\gamma$ and $\lambda$ are used to generate exponential decaying values, their goal is very different. The discount parameter $\gamma$ defines the objective of the MDP, the goal that we like to maximize. The exponential averaging parameter $\lambda$ is used to average over the different look-ahead parameters, and is selected to optimize the convergence.

The above describes the forward view of $TD(\lambda)$, where we average over future rewards. If we will try to implement it in a strict way this will lead us to wait until the end of the episode, since we will need to first observe all the rewards. Fortunately, there is an equivalent form of the $TD(\lambda)$ which uses a *backward view.* The backward view updates at each time step, using an incomplete information. At the end of the episode, the updates of the forward and backward updates will be the same.

The basic idea of the backward view is the following. Fix a time $\mathtt{t}$ and a state $\mathtt{s} = \mathtt{s_t}$. We have at time $\mathtt{t}$ a temporal difference $\Delta_\mathtt{t} = \mathtt{r_t} + \gamma V_\mathtt{t}(\mathtt{s_{t+1}}) - V_\mathtt{t}(\mathtt{s_t})$. Consider how this $\Delta_\mathtt{t}$ affects all the previous times $\tau < t$ where $\mathtt{s}_\tau = \mathtt{s} = \mathtt{s_t}$. The influence is exactly $(\gamma\lambda)^{t-\tau}\Delta_\mathtt{t}$. This implies that for every such $\tau$ we can do the desired update, however, we can aggregate all those updates to a single update. Let,

$$e_\mathtt{t}(\mathtt{s}) = \sum_{\tau \le t : \mathtt{s}_\tau = \mathtt{s}} (\gamma\lambda)^{t-\tau} = \sum_{\tau=1}^{\mathtt{t}} (\gamma\lambda)^{t-\tau} I(\mathtt{s}_\tau = \mathtt{s})$$

The above $e_\mathtt{t}(\mathtt{s})$ defines the *eligibility trace* and we can compute it online using

$$e_\mathtt{t}(\mathtt{s}) = \gamma\lambda e_{\mathtt{t}-1}(\mathtt{s}) + I(\mathtt{s} = \mathtt{s_t})$$

which result in the update

$$\widehat{V}_{\mathtt{t}+1}(\mathtt{s}) = \widehat{V}_\mathtt{t}(\mathtt{s}) + \alpha_\mathtt{t} e_\mathtt{t}(\mathtt{s})\Delta_\mathtt{t}$$

Note that for $TD(0)$ we have that $\lambda = 0$ and the eligibility trace becomes $e_\mathtt{t}(\mathtt{s}) = I(\mathtt{s} = \mathtt{s_t})$. This implies that we update only $\mathtt{s_t}$ and $\widehat{V}_{\mathtt{t}+1}(\mathtt{s_t}) = \widehat{V}_\mathtt{t}(\mathtt{s_t}) + \alpha_\mathtt{t}\Delta_\mathtt{t}$.

$TD(\lambda)$ **algorithm**

– Initialization: Set $\widehat{V}(\mathtt{s}) = 0$ (or any other value), and $e_0(\mathtt{s}) = 0$.
– Update: observe $(\mathtt{s_t}, \mathtt{a_t}, \mathtt{r_t}, \mathtt{s_{t+1}})$ and set

$$\begin{aligned} \Delta_\mathtt{t} &= \mathtt{r_t} + \gamma\widehat{V}_\mathtt{t}(\mathtt{s_{t+1}}) - \widehat{V}(\mathtt{s_t}) \\ e_\mathtt{t}(\mathtt{s}) &= \gamma\lambda e_{\mathtt{t}-1}(\mathtt{s}) + I(\mathtt{s} = \mathtt{s_t}) \\ \widehat{V}_{\mathtt{t}+1}(\mathtt{s}) &= \widehat{V}_\mathtt{t}(\mathtt{s}) + \alpha_\mathtt{t} e_\mathtt{t}(\mathtt{s})\Delta_\mathtt{t} \end{aligned}$$

To summarize, the benefit of $TD(\lambda)$ is that it interpolates between $TD(0)$ and Monte-carlo updates, and many times achieves superior performance to both. Similar to $TD(0)$, also $TD(\lambda)$ can be written as a stochastic approximation iterative algorithm, and one can derive its convergence. In the next section we show the equivalence of the forward and backward $TD(\lambda)$ updates.

### 9.4.3 The equivalence of the forward and backward view

We would like to show that indeed the forward and backward view result in the same overall update.

For the forward view we define the updates to be $\Delta_{\mathtt{t}}^F(\mathtt{s}) = \alpha(R_{\mathtt{t}}^\lambda - V_{\mathtt{t}}(\mathtt{s}))$, where $R_{\mathtt{t}}^\lambda = (1-\lambda)\sum_{n=1}^\infty \lambda^{n-1} R_{\mathtt{t}}^{(n)}$. Equivalently, $\Delta_{\mathtt{t}}^F(\mathtt{s}) = \alpha(1-\lambda)\sum_{n=1}^\infty \Delta_{\mathtt{t}}^{(n)}$,

For the backward view we define the updates to be $\Delta_{\mathtt{t}}^B(\mathtt{s}) = \alpha \Delta_{\mathtt{t}} e_{\mathtt{t}}(\mathtt{s})$, where the eligibility trace is $e_{\mathtt{t}}(\mathtt{s}) = \sum_{k=0}^{\mathtt{t}} (\lambda\gamma)^{t-k} I(\mathtt{s} = \mathtt{s}_k)$.

**Theorem 9.13.**

$$\sum_{\mathtt{t}=0}^\infty \Delta V_{\mathtt{t}}^B(\mathtt{s}) = \sum_{\mathtt{t}=0}^\infty \Delta V_{\mathtt{t}}^F(\mathtt{s}) I(\mathtt{s}_{\mathtt{t}} = \mathtt{s})$$

*Proof.* Consider the sum of the forward updates for state $\mathtt{s}$:

$$\begin{aligned}
\sum_{\mathtt{t}=0}^\infty \Delta \mathcal{V}_{\mathtt{t}}^F(\mathtt{s}) &= \sum_{\mathtt{t}=0}^\infty \alpha(1-\lambda) \sum_{n=\mathtt{t}}^\infty \lambda^{n-\mathtt{t}} \Delta_{\mathtt{t}}^{(n)} I(\mathtt{s} = \mathtt{s}_{\mathtt{t}}) \\
&= \sum_{\mathtt{t}=0}^\infty \alpha(1-\lambda) \sum_{n=\mathtt{t}}^\infty \lambda^{n-\mathtt{t}} \sum_{i=0}^n \gamma^i \Delta_{\mathtt{t}+i} I(\mathtt{s} = \mathtt{s}_{\mathtt{t}}) \\
&= \sum_{\mathtt{t}=0}^\infty \sum_{n=0}^\infty \sum_{k=\mathtt{t}}^n \alpha(1-\lambda)\lambda^{n-k}\lambda^{k-\mathtt{t}}\gamma^{k-\mathtt{t}} \Delta_k I(\mathtt{s} = \mathtt{s}_{\mathtt{t}}) \\
&= \sum_{\mathtt{t}=0}^\infty \sum_{k=\mathtt{t}}^\infty \alpha(\gamma\lambda)^{k-\mathtt{t}} \Delta_k I(\mathtt{s} = \mathtt{s}_{\mathtt{t}}) \sum_{i=0}^\infty (1-\lambda)\lambda^i \\
&= \sum_{\mathtt{t}=0}^\infty \sum_{k=\mathtt{t}}^\infty \alpha(\gamma\lambda)^{k-\mathtt{t}} \Delta_k(\mathtt{s}) I(\mathtt{s} = \mathtt{s}_{\mathtt{t}}) \qquad (9.1)
\end{aligned}$$

where the first identity is the definition, the second identity follows since $\Delta_{\mathtt{t}}^{(n)} = \sum_{i=0}^n \gamma^i \Delta_{\mathtt{t}+i}$, in the third identity we substitute $k$ for $t+i$ and sum over $n$, $k$ and $\mathtt{t}$, in the forth identity we substitute $i$ for $n-k$ and isolate the terms that depend on $i$, and in the last identity we note that $\sum_{i=0}^\infty (1-\lambda)\lambda^i = 1$.

For the backward view for state $\mathtt{s}$ we have

$$\begin{aligned}
\sum_{\mathtt{t}=0}^\infty \Delta \mathcal{V}_{\mathtt{t}}^B(\mathtt{s}) &= \sum_{\mathtt{t}=0}^\infty \alpha \Delta_{\mathtt{t}}(\mathtt{s}) \sum_{k=0}^{\mathtt{t}} (\gamma\lambda)^{\mathtt{t}-k} I(\mathtt{s} = \mathtt{s}_{\mathtt{t}}) \\
&= \sum_{k=0}^\infty \sum_{\mathtt{t}=k}^\infty \alpha(\gamma\lambda)^{\mathtt{t}-k} \Delta_{\mathtt{t}}(\mathtt{s}) I(\mathtt{s} = \mathtt{s}_{\mathtt{t}}) \qquad (9.2)
\end{aligned}$$

Note that if we interchange $k$ and $\mathtt{t}$ in Eq. (9.1) and in Eq. (9.2), then we have the identical expressions. □

### 9.4.4 *SARSA*($\lambda$)

We can use the idea of eligibility traces also in other algorithms, such as SARSA. Recall that given $(\mathtt{s_t}, \mathtt{a_t}, \mathtt{r_t}, \mathtt{s_{t+1}}, \mathtt{a_{t+1}})$ the update of SARSA is

$$\mathtt{r_t} + \gamma Q_\mathtt{t}(\mathtt{s_{t+1}}, \mathtt{a_{t+1}}) - Q_\mathtt{t}(\mathtt{s_t}, \mathtt{a_t})$$

Similarly, we can define an $n$-step look-ahead $q_\mathtt{t}^{(n)} = \sum_{i=0}^{n-1} \gamma^i \mathtt{r}_{\mathtt{t}+i} + \gamma^n Q_\mathtt{t}(\mathtt{s}_{\mathtt{t}+n}, \mathtt{a}_{\mathtt{t}+n})$ and set $Q_{\mathtt{t+1}}(\mathtt{s_t}, \mathtt{a_t}) = Q_\mathtt{t}(\mathtt{s_t}, \mathtt{a_t}) + \alpha_\mathtt{t}(q_\mathtt{t}^{(n)} - Q_\mathtt{t}(\mathtt{s_t}, \mathtt{a_t}))$.

We can now define *SARSA*($\lambda$) using exponential averaging with parameter $\lambda$. Namely, we define $q_\mathtt{t}^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} q_\mathtt{t}^{(n)}$. This makes the forward view of *SARSA*($\lambda$) to be $Q_{\mathtt{t+1}}(\mathtt{s_t}, \mathtt{a_t}) = Q_\mathtt{t}(\mathtt{s_t}, \mathtt{a_t}) + \alpha_\mathtt{t}(q_\mathtt{t}^\lambda - Q_\mathtt{t}(\mathtt{s_t}, \mathtt{a_t}))$.

Similar to $TD(\lambda)$, we can define a backward view using eligibility traces:

$$\begin{aligned} &e_0(\mathtt{s}, \mathtt{a}) = 0 \\ &e_\mathtt{t}(\mathtt{s}, \mathtt{a}) = \gamma\lambda e_{\mathtt{t-1}}(\mathtt{s}, \mathtt{a}) + I(\mathtt{s} = \mathtt{s_t}, a = \mathtt{a_t}) \end{aligned}$$

For the update we have

$$\begin{aligned} \Delta_\mathtt{t} &= \mathtt{r_t} + \gamma Q_\mathtt{t}(\mathtt{s_{t+1}}, \mathtt{a_{t+1}}) - Q_\mathtt{t}(\mathtt{s_t}, \mathtt{a_t}) \\ Q_{\mathtt{t+1}}(\mathtt{s}, \mathtt{a}) &= Q_\mathtt{t}(\mathtt{s}, \mathtt{a}) + \alpha_\mathtt{t} e_\mathtt{t}(\mathtt{s}, \mathtt{a})\Delta_\mathtt{t} \end{aligned}$$

## 9.5 Miscellaneous

### 9.5.1 Importance Sampling

Importance sapling is a simple general technique to estimate the mean with respect to a given distribution, while sampling from a different distribution. To be specific, let $Q$ be the sampling distribution and $P$ the evaluation distribution. The basic idea is the following

$$E_{x\sim P}[f(x)] = \sum_x P(x)f(x) = \sum_x Q(x)\frac{P(x)}{Q(x)}f(x) = E_{x\sim Q}[\frac{P(x)}{Q(x)}f(x)]$$

This implies that given a sample $\{x_1, \ldots, x_m\}$ from $Q$, we can estimate $E_{x\sim P}[f(x)]$ using $\sum_{i=1}^m \frac{P(x_i)}{Q(x_i)} f(x_i)$.

The importance sampling gives an unbiased estimator, but the variance of the estimator might be huge, since it depends on $P(x)/Q(x)$.

We would like to apply the idea of importance sampling to learning in MDPs. Assume that there is a policy $\pi$ that selects the actions, and there is a policy $\rho$ that we would like to evaluate. For the importance sampling, given a trajectory, we need to take the ratio of probabilities under $\rho$ and $\pi$.

$$\frac{\rho(\mathtt{s}_1, \mathtt{a}_1, \mathtt{r}_1, \ldots, \mathtt{s}_T, \mathtt{a}_T, \mathtt{r}_T, \mathtt{s}_{T+1})}{\pi(\mathtt{s}_1, \mathtt{a}_1, \mathtt{r}_1, \ldots, \mathtt{s}_T, \mathtt{a}_T, \mathtt{r}_T, \mathtt{s}_{T+1})} = \prod_{\mathtt{t}=1}^{T} \frac{\rho(a_\mathtt{t}|\mathtt{s}_\mathtt{t})}{\pi(a_\mathtt{t}|\mathtt{s}_\mathtt{t})}$$

where the equality follows since the reward and transition probabilities are identical, and cancel.

For Monte-Carlo, the estimates would be

$$G^{\rho/\pi} = \prod_{\mathtt{t}=1}^{T} \frac{\rho(\mathtt{a}_\mathtt{t}|\mathtt{s}_\mathtt{t})}{\pi(\mathtt{a}_\mathtt{t}|\mathtt{s}_\mathtt{t})} (\sum_{\mathtt{t}=1}^{T} \mathtt{r}_\mathtt{t})$$

and we have

$$\widehat{V}^\rho(\mathtt{s}_1) = \widehat{V}^\rho(\mathtt{s}_1) + \alpha(G^{\rho/\pi} - \widehat{V}^\rho(\mathtt{s}_1))$$

This updates might be huge, since we are multiplying the ratios of many small numbers.

For the $TD(0)$ the updates will be

$$\Delta_\mathtt{t}^{\rho/\pi} = \frac{\rho(\mathtt{a}_\mathtt{t}|\mathtt{s}_\mathtt{t})}{\pi(\mathtt{a}_\mathtt{t}|\mathtt{s}_\mathtt{t})} \mathtt{r}_\mathtt{t} + \gamma \widehat{V}(\mathtt{s}_{\mathtt{t}+1}) - \widehat{V}(\mathtt{s}_\mathtt{t})$$

and we have

$$\widehat{V}^\rho(\mathtt{s}_1) = \widehat{V}^\rho(\mathtt{s}_1) + \alpha(\Delta_\mathtt{t}^{\rho/\pi} - \widehat{V}^\rho(\mathtt{s}_1))$$

This update is much more stable, since we have only one factor multiplying the observed reward.

### 9.5.2 Actor-critic methodology

Actor-critic gives a general methodology of building reinforcement learning algorithms. It is composed from an actor, that selects the actions, and a critic, that learns the value function. The actor observes the current state, and the value function, and selects an action. The critic, observes the current state (and action) and reward and outputs a value function. See Figure 9.5.

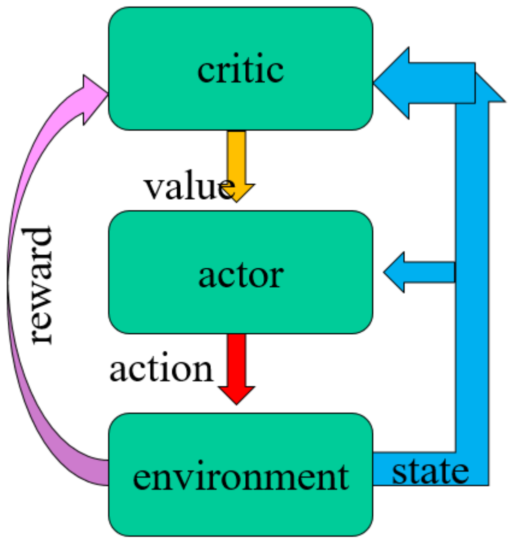


Figure 9.5: The situated agent

## 9.6 Bibliography Remarks

The $Q$-learning outline of the asymptotic convergence and the step size analysis follows [4].

Expected SARSA was presented in [18]

The comparison of the `First Visit` and `Every Visit` is based on [12].

Part of the outline borrows from David Silver class notes and the the book of Sutton and Barto [14].

The presentation of the TD algorithms follows the book of Sutton and Barto [14].

# Chapter 10

# Tabular learning: off-policy

# Chapter 11

# Large state space: value function approximation

This chapter starts looking at the case where the MDP models are large. In the current chapter we will look at approximating the value function. In the next chapter we will consider learning directly a policy and optimizing it.

When we talk about a large MDP, it can be in one of a few different reasons. The most common is having a large state space. For example, Backgammon has $10^{20}$ states, Go has $10^{170}$ and robot control has continuous state space. Another dimension is the action space, which can be even continuous in many applications (say, robots). Finally, we might have a complex dynamics which are hard to describe succinctly. The main challenge is that we need our algorithm to scale up to this enormous spaces.

Previously, we had a look-up table for the value function $\mathcal{V}^{\pi}(\mathtt{s})$ or $Q^{\pi}(\mathtt{s},\mathtt{a})$, which we updated every time we encountered the state $\mathtt{s}$. Here, we will consider function approximation for large MDPs, mainly large state spaces. In a large state space this will be infeasible. Not only that we do not have the memory, but even more importantly, we are unlikely to observe states re-occurring. We will need to make (implicit) assumptions about the MDP, which can be viewed similarly to our assumption in learning a classifier in supervised learning.

Specifically, we will have the value functions parameterized by *weights* $\mathtt{w}$, namely, $\widehat{V}(\mathtt{s};\mathtt{w}) \approx \mathcal{V}^{\pi}(\mathtt{s})$ and $\widehat{Q}(\mathtt{s},\mathtt{a};\mathtt{w}) \approx Q^{\pi}(\mathtt{s},\mathtt{a})$. We would like to guarantee generalization from the observed states and trajectories to unobserved states and trajectories. In the process we will update the weights $\mathtt{w}$, using some learning procedure, most notably using MC or TD learning.

When considering a value function there are a few interpretations what exactly we mean. It can either be: (1) mapping from a state $\mathtt{s}$ to its expected return, i.e., $\widehat{V}^\pi(\mathtt{s};\mathtt{w})$. (2) mapping from state-action pairs $(\mathtt{s},\mathtt{a})$ to their expected return, i.e., $\widehat{Q}^\pi(\mathtt{s},\mathtt{a};\mathtt{w})$. (3) mapping from states $\mathtt{s}$ to expected return of each action, i.e., $\langle \widehat{Q}^\pi(\mathtt{s},\mathtt{a}_i;\mathtt{w}) : \mathtt{a}_i \in \mathcal{A}\rangle$. All the interpretations are valid, and our discussion will not distinguish between them. (Actually, for $\langle \widehat{Q}^\pi(\mathtt{s},\mathtt{a}_i;\mathtt{w}) : \mathtt{a}_i \in \mathcal{A}\rangle$ we implicitly assume that the number of actions is small.)

We now need to discuss how will we build the approximating function. For this we can turn to the rich literature in Machine Learning and consider popular hypothesis classes. For example: (1) Linear functions, (2) Neural networks, (3) Decision trees, (4) Nearest neighbors, (5) Fourier or wavelet basis, etc. We will concentrate here on linear functions and neural networks, mainly since gradient based methods apply to them very naturally.

We will consider in this chapter (mainly) the discounted return with a discount parameter $\gamma \in (0,1)$. The results extend very naturally to the finite horizon and episodic settings.

**YM: Maybe better to switch to finite horizon, and avoid the need for mixing time and stead state distribution?**

## 11.1 Basic Approximation

Before we start the discussion on the learning methods, we will do a small detour. We will discuss the effect of having an error in the value function we learn, and its effect on the outcome. Assume we have a value function $V$ such that $\|V - \mathcal{V}^*\|_\infty \leq \varepsilon$. Let $\pi$ be the greedy policy with respect to $V$, namely,

$$\pi(\mathtt{s}) = \arg\max_{\mathtt{a}}[\mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma E_{\mathtt{s}'\sim p(\cdot|\mathtt{s},\mathtt{a})}[V(\mathtt{s}')]$$

**Theorem 11.1.** *Let $V$ such that $\|V - \mathcal{V}^*\|_\infty \leq \varepsilon$ and $\pi$ be the greedy policy with respect to $V$. Then,*

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

*Proof.* Consider two operators $\mathcal{T}^\pi$ and $\mathcal{T}^*$ (see Chapter 5.4.3). The first $\mathcal{T}^\pi$ is

$$(\mathcal{T}^\pi v)(\mathtt{s}) = \mathtt{r}(\mathtt{s},\pi(\mathtt{s})) + \gamma E_{\mathtt{s}'\sim p(\cdot|\mathtt{s},\pi(\mathtt{s}))}[v(\mathtt{s}')],$$

and it converges to $\mathcal{V}^\pi$ (see Theorem 5.9).

The second $T^*$ is

$$(\mathcal{T}^* v)(\mathtt{s}) = \max_{\mathtt{a}}[\mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma E_{\mathtt{s}'\sim p(\cdot|\mathtt{s},\mathtt{a})}[v(\mathtt{s}')]],$$

and it converges to $\mathcal{V}^*$ (see Theorem 5.9).

In addition, recall that we have showed that both $\mathcal{T}^\pi$ and $\mathcal{T}^*$ are $\gamma$-contracting (see Theorem 5.9).

Since $\pi$ is greedy with respect to $V$ we have $\mathcal{T}^\pi V = \mathcal{T}^* V$ (but this does not hold for other value functions $V'$).

Then,

$$\begin{aligned}
\|\mathcal{V}^\pi - \mathcal{V}^*\|_\infty &= \|T^\pi \mathcal{V}^\pi - \mathcal{V}^*\|_\infty \\
&\leq \|T^\pi \mathcal{V}^\pi - T^\pi V\|_\infty + \|T^\pi V - \mathcal{V}^*\|_\infty \\
&\leq \gamma \|\mathcal{V}^\pi - V\|_\infty + \|T^* V - T^* \mathcal{V}^*\|_\infty \\
&\leq \gamma \|\mathcal{V}^\pi - V\|_\infty + \gamma \|V - \mathcal{V}^*\|_\infty \\
&\leq \gamma(\|\mathcal{V}^\pi - \mathcal{V}^*\|_\infty + \|\mathcal{V}^* - \mathcal{V}\|_\infty) + \gamma \|V - \mathcal{V}^*\|_\infty
\end{aligned}$$

where in the second inequality we used the fact that since since $\pi$ is greedy with respect to $V$ then $T^\pi V = T^* V$.

Reorganizing the inequality and recalling that $\|\mathcal{V}^* - \mathcal{V}\|_\infty \leq \varepsilon$, we have

$$(1-\gamma)\|\mathcal{V}^\pi - \mathcal{V}^*\|_\infty \leq 2\varepsilon\gamma$$

and the theorem follows. □

The above theorem states that if we have a small errors in $L_\infty$ norm, the effect of the errors on the expected return is bounded. However, in most cases we will not be able to guarantee an approximation in nor $L_\infty$. This is since it is infeasible even to compute the $L_\infty$ norm of two given value function, since this requires considering *all* states. In the large state space setting such operations are infeasible.

## 11.2 From RL to ML

We would like to reduce our reinforcement learning problem to a supervised learning problem. This will enable us to use any of the many techniques of machine learning to address the problem. Let us consider the basic ingredients of supervised learning. The most important ingredient is having a labeled sample set, which is sampled i.i.d.

Let us start by considering an idealized setting. Fix a policy $\pi$ that we like to learn its value function. We like to generate a sample

$$\{(\mathtt{s}_1, \mathcal{V}^\pi(\mathtt{s}_1)), \ldots, (\mathtt{s}_m, \mathcal{V}^\pi(\mathtt{s}_m))\}$$

We first need to discuss how to sample the states $\mathtt{s}_i$ in an i.i.d. way. We can generate trajectory, but we need to be careful, since adjacent states are definitely dependent! One solution is to space the sampling from the trajectory using the mixing time of $\pi$.[1] This will give us samples $\mathtt{s}_i$ which are (almost) from the stationary distribution of $\pi$ and are (almost) independent.

The hardest, and most confusing, ingredient is the labels $\mathcal{V}^{\pi}(\mathtt{s}_i)$. In machine leaning we actually assume that someone gives us the labels to build a classifier. If someone is able to give us $\mathcal{V}^{\pi}(\mathtt{s})$, it seems like assuming the problem away, since this is what we would like to learn.

Actually we need to define three things: (1) states and actions, which we will do using $\pi$, (2) a loss function, which will give the tradeoffs between different approximation errors, and (3) labels, which we are not given explicitly.

For the state and actions we will define a threshold $\tau$, which will depends on the mixing time. We will run the policy $\pi$ for $\tau$ steps and then output the state $\mathtt{s}$ (and action $\mathtt{a}$, if needed).

For the loss function we will define a differential loss function $J(\mathtt{w})$. This will enable us to compute the gradient of the loss function $\nabla_{\mathtt{w}} J(\mathtt{w}) = \langle \frac{\partial}{\partial \mathtt{w}_1} J(\mathtt{w}), \ldots, \frac{\partial}{\partial \mathtt{w}_d} J(\mathtt{w}) \rangle$, and update the weights in the negative direction of the gradient, namely, $\Delta \mathtt{w} = -\alpha \nabla_{\mathtt{w}} J(\mathtt{w})$, where $\alpha$ is the step size of the learning rate. In supervised learning we are guaranteed that the process will converge to a local minimum (or a saddle point).

Let us start by defining the loss function,

$$J(\mathtt{w}) = \frac{1}{2} \sum_{s} \mu(\mathtt{s}) (\mathcal{V}^{\pi}(\mathtt{s}) - \widehat{V}^{\pi}(\mathtt{s}; \mathtt{w}))^2$$

where $\mu(\mathtt{s})$ is the steady state distribution of $\pi$.[2] So we can compute the gradient and update the weights. The Stochastic Gradient Descent (SGD) is simply sampling a single state $\mathtt{s}$ and using it to update, i.e.,

$$\Delta \mathtt{w}_{\mathtt{t}} = \alpha (\mathcal{V}^{\pi}(\mathtt{s}) - \widehat{V}^{\pi}(\mathtt{s}; \mathtt{w})) \nabla \widehat{V}^{\pi}(\mathtt{s}; \mathtt{w})$$

We now need to build the sample, namely, how do we set the labels to replace $\mathcal{V}^{\pi}(\mathtt{s})$. The basic idea is to find an unbiased estimator $U_{\mathtt{t}}$ such that $E[U_{\mathtt{t}} | \mathtt{s}_{\mathtt{t}}] = \mathcal{V}^{\pi}(\mathtt{s}_{\mathtt{t}})$. We now will present a few different approaches to derive such unbiased estimators.

The simplest approach is to use Monte-Carlo (MC) sampling.[3] Recall that in First Visit Monte-Carlo we have $R_{\mathtt{t}}(\mathtt{s}) = \sum_{\tau=1}^{T} \mathtt{r}_{\tau}$, starting at the first visit of $\mathtt{s}$ in

[1]See Chapter 3 for definition.
[2]See Chapter 3 for definition.
[3]See Chapter 9.3

episode $\mathtt{t}$. Clearly, we have $E[R_{\mathtt{t}}(\mathtt{s})] = \mathcal{V}^{\pi}(\mathtt{s})$, since samples are independent, so we can set $U_{\mathtt{t}}(\mathtt{s}) = R_{\mathtt{t}}(\mathtt{s})$. The update becomes,

$$\Delta \mathtt{w}_{\mathtt{t}} = \alpha(R_{\mathtt{t}}(\mathtt{s}) - \widehat{V}^{\pi}(\mathtt{s};\mathtt{w}_{\mathtt{t}}))\nabla_{w}\widehat{V}^{\pi}(\mathtt{s};\mathtt{w})$$

We can try to use the same idea for $TD(0)$.[4] The estimate of $TD(0)$ for $\mathtt{s}_{\mathtt{t}}$ is $R_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}}) = \mathtt{r}_{\mathtt{t}} + \gamma\widehat{V}^{\pi}(\mathtt{s}_{\mathtt{t}+1};\mathtt{w})$. The first problem is that this is a biased estimator since

$$\mathcal{V}^{\pi}(\mathtt{s}_{\mathtt{t}}) = E[\mathtt{r}_{\mathtt{t}} + \gamma\mathcal{V}^{\pi}(\mathtt{s}_{\mathtt{t}+1})] \neq E[\mathtt{r}_{\mathtt{t}} + \gamma\widehat{V}(\mathtt{s}_{\mathtt{t}+1};\mathtt{w})]$$

We have an additional problem with the gradient, since we have $\mathtt{w}$ influencing also the target $R_{\mathtt{t}}(\mathtt{s}_{\mathtt{t}})$ through $\widehat{V}^{\pi}(\mathtt{s}_{\mathtt{t}+1};\mathtt{w})$, something we do not have in MC (or generally in ML). For this reason $\nabla_{\mathtt{w}}\widehat{V}(\mathtt{s}_{\mathtt{t}};\mathtt{w})$ is called a semi-gradient. The update becomes,

$$\Delta \mathtt{w}(\mathtt{s}) = \alpha[\mathtt{r}_{\mathtt{t}} + \gamma\widehat{V}(\mathtt{s}_{\mathtt{t}+1};\mathtt{w}) - \widehat{V}(\mathtt{s}_{\mathtt{t}})]\nabla_{\mathtt{w}}\widehat{V}(\mathtt{s}_{\mathtt{t}};\mathtt{w})$$

Finally we get to $TD(\lambda)$. Similar to $TD(0)$ we can now define the forward update to be,

$$\Delta \mathtt{w} = \alpha[R_{\mathtt{t}}^{\lambda} - \widehat{V}(\mathtt{s}_{\mathtt{t}};\mathtt{w})]\nabla_{\mathtt{w}}\widehat{V}(\mathtt{s}_{\mathtt{t}};\mathtt{w})$$

and the backward update,

$$\begin{aligned} e_{\mathtt{t}} &= \gamma\lambda e_{\mathtt{t}-1} + \nabla_{\mathtt{w}}\widehat{V}(\mathtt{s}_{\mathtt{t}};\mathtt{w}) \\ \Delta_{\mathtt{t}} &= \mathtt{r}_{\mathtt{t}} + \gamma\widehat{V}(\mathtt{s}_{\mathtt{t}+1};\mathtt{w}) - \widehat{V}(\mathtt{s}_{t};\mathtt{w}) \\ \Delta \mathtt{w} &= \alpha\Delta_{\mathtt{t}}e_{\mathtt{t}} \end{aligned}$$

## 11.3 Linear Functions

We first start with the state encoding. In general we assume that each state $\mathtt{s}$ is encoded by a vector $x(\mathtt{s}) \in \mathbb{R}^{d}$. For notation, let $x(\mathtt{s}) = (x_1(\mathtt{s}), \ldots, x_d(\mathtt{s}))$. This will be useful for any hypothesis, and specifically, linear function and neural networks.

A linear function is characterized by a vector of weights $\mathtt{w} \in \mathbb{R}^{d}$. The value of the function at state $\mathtt{s}$ is

$$\widehat{V}(\mathtt{s};\mathtt{w}) = \mathtt{w}^{\top}x(\mathtt{s})$$

This implies that the SGD updates become

$$\mathtt{w}_{\mathtt{t}+1} = \mathtt{w}_{\mathtt{t}}\alpha[U_{\mathtt{t}} - \widehat{V}(\mathtt{s}_{\mathtt{t}};\mathtt{w}_{\mathtt{t}})]x(\mathtt{s}_{\mathtt{t}})$$

[4]See Chapter 9.4.

The main benefit of the linear function is when $d \ll |\mathcal{S}|$. When $d = |\mathcal{S}|$ we can simply encode a look-up table using $\mathtt{w}$. This is done by setting the encoding of each state to be the unit vector $x_i(\mathtt{s}) = I(\mathtt{s} = \mathtt{s}_i)$, and the approximation to be $\widehat{V}(\mathtt{s};\mathtt{w}) = \sum_{i=1}^{d} \mathtt{w}_i I(\mathtt{s} = \mathtt{s}_i)$. This implies that $\widehat{V}(\mathtt{s}_i;\mathtt{w}) = \mathtt{w}_i$.

For a linear function the gradient is simply $x(\mathtt{s})$, i.e., $\nabla_{\mathtt{w}} \mathtt{w}^\top x(\mathtt{s}) = x(\mathtt{s})$. The Monte-Carlo update will become

$$\Delta\mathtt{w} = \alpha[R_{\mathtt{t}} - \widehat{V}(\mathtt{s}_{\mathtt{t}};\mathtt{w}_{\mathtt{t}})]x(\mathtt{s}_{\mathtt{t}})$$

For $TD(0)$ we have

$$\Delta\mathtt{w} = \alpha[\mathtt{r}_{\mathtt{t}} + \gamma\widehat{V}(\mathtt{s}_{\mathtt{t}+1};\mathtt{w}_{\mathtt{t}}) - \widehat{V}(\mathtt{s}_{\mathtt{t}};\mathtt{w}_{\mathtt{t}})]x(\mathtt{s}_{\mathtt{t}})$$

For $TD(\lambda)$, for the forward view we have

$$\Delta\mathtt{w} = \alpha[R^{\lambda}_{\mathtt{t}} - \widehat{V}(\mathtt{s}_{\mathtt{t}};\mathtt{w}_{\mathtt{t}})]x(\mathtt{s}_{\mathtt{t}})$$

For the backward view we have,

$$\begin{aligned} e_{\mathtt{t}} &= \gamma\lambda e_{\mathtt{t}-1} + x(\mathtt{s}_{\mathtt{t}}) \\ \Delta_{\mathtt{t}} &= \mathtt{r}_{\mathtt{t}} + \gamma\widehat{V}(\mathtt{s}_{\mathtt{t}+1};\mathtt{w}_{\mathtt{t}}) - \widehat{V}(\mathtt{s}_{t};\mathtt{w}_{\mathtt{t}}) \\ \Delta\mathtt{w} &= \alpha\Delta_{\mathtt{t}} e_{\mathtt{t}} \end{aligned}$$

We would like to discuss the convergence of the different algorithms. Recall that the mean squared error loss function is strongly convex, i.e.,

$$J(\mathtt{w}) = \frac{1}{2}\sum_{s} \mu(\mathtt{s})(\mathcal{V}^{\pi}(\mathtt{s}) - \widehat{V}^{\pi}(\mathtt{s};\mathtt{w}))^2$$

Therefore there is a unique $\mathtt{w}_{min}$ which minimizes $J(\mathtt{w})$. Monte-Carlo is simply running a Stochastic Gradient Decent (SGD) algorithm, and therefore the convergence and its rate follow from known results in convex optimization.

The convergence of $TD$ is more problematic, since it uses a semi-gradient and not the true gradient. The good news is that for linear functions, when using an on-policy, there is a proof of convergence. However, the convergence is not to $\mathtt{w}_{min}$ but rather to a different point $\mathtt{w}_{TD}$. The difference is due to the fact that we are minimizing the expression

$$\mathtt{w}_{TD} = \arg\min_{\mathtt{w}} \frac{1}{2}\sum_{s} \mu(\mathtt{s}) \left(\mathtt{r}(\mathtt{s},\pi(\mathtt{s})) + \gamma E_{\mathtt{s}'}[\widehat{V}(\mathtt{s}';\mathtt{w}_{\mathtt{t}})] - \widehat{V}^{\pi}(\mathtt{s};\mathtt{w})\right)^2$$

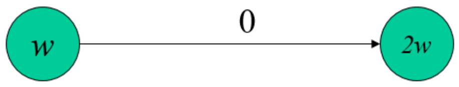


Figure 11.1: Two state snippet of an MDP

The difference in the losses can be bounded as follows[5]

$$J(\mathtt{w}_{TD}) \leq \frac{1}{1-\gamma} J(\mathtt{w}_{min})$$

Note that for $\gamma \approx 1$, which is a popular setting, we might have a huge multiplicative factor, yet bounded. However, if we are in the realizable case, $J(\mathtt{w}_{min}) = 0$, then we have also $J(\mathtt{w}_{TD}) = 0$.

## 11.4 Off-policy

We would like to see what is the effect that the samples come when following a different policy, namely, an off-policy setting. There is no issue for Monte-Carlo, and the same logic would still be valid. For TD, we did not have any problem in the look-up model. We would like to see what can go wrong when we have a function approximation setting.

Consider the following part of an MDP (see Figure 11.1) consists of two nodes, with a transition from the first to the second, with reward 0. The main issue is that the linear approximation gives the first node a weight $\mathtt{w}$ and the second $2\mathtt{w}$. Assume we start with some value $\mathtt{w}_0 > 0$. Each time we have an update for the two states we have

$$\mathtt{w}_{\mathtt{t}+1} = \mathtt{w}_{\mathtt{t}} + \alpha[0 + \gamma(2\mathtt{w}_{\mathtt{t}}) - \mathtt{w}_{\mathtt{t}}] = [1 + \alpha(2\gamma - 1)]\mathtt{w}_{\mathtt{t}}$$

For $\gamma > 0.5$ we have $\alpha(2\gamma - 1) > 0$, and $\mathtt{w}_{\mathtt{t}}$ diverges.

We are implicitly assuming that the setting is off-policy, since in an on-policy, we would continue from the second state, and eventually lower the weight.

To have a "complete" example consider the three state MDP in Figure 11.2. All the rewards are zero, and the main difference is that we have a new terminating state, that we reach with probability $p$.

[5] The derivation appears in Section 11.6.

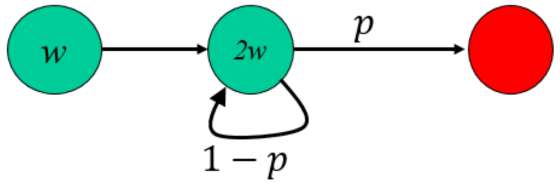


Figure 11.2: The three state MDP. All rewards are zero.

Again, assume that we start with some $\mathtt{w}_0 > 0$ We have three types of updates, one per possible transition. When we transition from the initial state to the second state we have

$$\Delta\mathtt{w} = \alpha[0 + \gamma(2\mathtt{w_t}) - \mathtt{w_t}] \cdot 1 = \alpha(2\gamma - 1)\mathtt{w_t}$$

The transition from the second state back to itself has an update,

$$\Delta\mathtt{w} = \alpha[0 + \gamma(2\mathtt{w_t}) - (2\mathtt{w_t})] \cdot 2 = -4\alpha(1-\gamma)\mathtt{w_t}$$

The transition to the terminal state we have

$$\Delta\mathtt{w} = \alpha[0 + \gamma 0 - (2\mathtt{w_t})] \cdot 2 = -4\alpha\mathtt{w_t}$$

When we use on-policy, we have all transitions. Assume that the second transition happens $n \geq 0$ times. Then we have

$$\frac{\mathtt{w_{t+1}}}{\mathtt{w_t}} = (1 + \alpha(2\gamma - 1))(1 - 4\alpha(1-\gamma))^n(1 - 4\alpha) < 1 - \alpha$$

This implies that $\mathtt{w_t}$ converges to zero, as desired.

Now consider an off-policy that truncates the episodes after $n$ transitions of the second state, where $n \ll 1/p$, and in addition $\gamma > 1 - 1/(40n)$. This implies that in most updates we do not reach the terminal state and we have

$$\frac{\mathtt{w_{t+1}}}{\mathtt{w_t}} = (1 + \alpha(2\gamma - 1))(1 - 4\alpha(1-\gamma))^n > 1$$

and therefore, for the some setting of $n$ we have that the weight $\mathtt{w_t}$ diverges.

We might hope that the divergence is due to the online nature of the TD updates. We can consider an algorithm that in each iteration minimizes the square error. Namely,

$$\mathtt{w_{t+1}} = \arg\min_{\mathtt{w}} \sum_s [\widehat{V}(\mathtt{s_t}; \mathtt{w}) - E^\pi[\mathtt{r_t} + \gamma\widehat{V}(\mathtt{s_{t+1}}; \mathtt{w_t})]]^2$$

For the MDP example of Figure 11.2 we have that:

$$\mathtt{w_{t+1}} = \arg\min_{\mathtt{w}} (\mathtt{w} - \gamma(2\mathtt{w_t}))^2 + (2\mathtt{w} - (1-p)\gamma(2\mathtt{w_t}))^2$$

Solving for $\mathtt{w_{t+1}}$ we have

$$\begin{aligned} 0 &= 2(\mathtt{w_{t+1}} - \gamma(2\mathtt{w_t})) + 4(2\mathtt{w_{t+1}} - (1-p)\gamma(2\mathtt{w_t}) \\ 10\mathtt{w_{t+1}} &= 4\gamma\mathtt{w_t} + 8\gamma(1-p)\mathtt{w_t} \\ \mathtt{w_{t+1}} &= \frac{6-4p}{5}\gamma\mathtt{w_t} \end{aligned}$$

So for $\gamma(6-4p) > 5$ we have divergence. (Recall that $\gamma \approx 1-\varepsilon$ and $p \approx \varepsilon$ is a very important setting.)

Note that if we have taken in to account the influence of $\mathtt{w_{t+1}}$ on $V$ and use $\widehat{V}(\mathtt{s_{t+1}}; \mathtt{w_{t+1}})$ instead of $\widehat{V}(\mathtt{s_{t+1}}; \mathtt{w_t})$, this specific problem would have disappeared, since $\mathtt{w_{t+1}} = 0$ would be the minimizer.

## 11.5 Summary

What we have seen, as Rich Sutton says, that there is a *deadly triad*:

1. Function Approximation
2. Bootstrapping methods, such as TD.
3. Off-policy training.

When the three conditions hold, we man times have counter examples for the convergence.

Here is a summary of the known convergence and divergence results in the literature:

| | algorithm | look-up table | linear function | non-linear |
|---|---|---|---|---|
| on-policy | MC | + | + | + |
| on-policy | $TD(0), TD(\lambda)$ | + | + | - |
| off-policy | MC | + | + | + |
| off-policy | $TD(0), TD(\lambda)$ | + | - | - |

The results for the look-up table where derived in Chapter 9. The fact that Monte-Carlo methods converge is due to the fact that they are running an SGD algorithm.

For linear functions with convex loss they will converge to the global optimum and for non-linear functions (for example, neural networks) they will converge to a local optima. The convergence to the TD appear in Chapter 11.6. The divergence of TD with linear fucntions in an off-policy setting appear in Chapter 11.4. The TD divergence in the non-linear online setting appears in [17].

## 11.6 Advanced topics

### 11.6.1 Comparing $\mathtt{w}_{min}$ and $\mathtt{w}_{TD}$

The derivation follows [17], however, for simplicity, we limit ourselves to $TD(0)$ rather than $TD(\lambda)$.

We will need to start with a few definitions. Given the steady state distribution $\mu(\mathtt{s})$ for $\mathtt{s} \in \mathcal{S}$ we define a diagonal matrix $D$ of size $|\mathcal{S}| \times |\mathcal{S}|$ where $D(\mathtt{s}, \mathtt{s}) = \mu(\mathtt{s})$. We define an inner-product $\langle \cdot, \cdot \rangle_D$ where $\langle z_1, z_2 \rangle_D = z_1^\top D z_2$ and a norm $\|z\|_D = \sqrt{< z, z >_D}$.

We define $\mathcal{T}^\pi(V) = \mathtt{r} + \gamma P V$ and recall that: (1) $\mathcal{V}^\pi = \mathcal{T}^\pi(\mathcal{V}^\pi)$, (2) $\|\mathcal{T}^\pi(V_1) - \mathcal{T}^\pi(V_2)\|_D \leq \gamma \|V_1 - V_2\|_D$ (actually we we need to prove that for $\|\cdot\|_D$).

We will perform a linear function approximation where the attributes of state $\mathtt{s}$ are $x_\mathtt{s}$. Define a matrix $X$ of size $|\mathcal{S}| \times d$, where the $\mathtt{s}$ row is $x_\mathtt{s}$. Then using a parameter $\mathtt{w}$ our approximation is $\widehat{V}(\cdot; \mathtt{w}) = X\mathtt{w}$. We now need to project vectors to an approximation $X\mathtt{w}$. Give a vector $V$, let $\Pi(V) = \arg\min_\mathtt{w} \|X\mathtt{w} - V\|_D$. This projection operation can be written explicitly as

$$\Pi(V) = X(X^\top D X)^{-1} X^\top D V.$$

We can write the expected update as:

$$\Delta \mathtt{w} = X^\top D(T^\pi(X\mathtt{w}_\mathtt{t}) - X\mathtt{w})$$

which is the gradient of

$$\nabla_\mathtt{w} \|T^\pi(X\mathtt{w}_\mathtt{t}) - Xw\|_D^2$$

note that the input to $T^\pi$ is fixed and we optimize over $\mathtt{w}$ only. The minimization is

$$\mathtt{w}_{\mathtt{t+1}} = \Pi(T^\pi(X\mathtt{w}_\mathtt{t}))$$

We will need the following technical claim:

**Claim 11.2.**

$$\|PV\|_D^2 \leq \|V\|_D^2$$

*Proof.*

$$\begin{aligned}
\|PV\|_D^2 &= V^\top P^\top DPV \\
&= \sum_s \mu(\mathtt{s})(\sum_{\mathtt{s}'} p(\mathtt{s}'|\mathtt{s})V(\mathtt{s}'))^2 \\
&\le \sum_s \mu(\mathtt{s}) \sum_{\mathtt{s}'} p(\mathtt{s}'|\mathtt{s})V^2(\mathtt{s}') \\
&= \sum_{\mathtt{s}'} V^2(\mathtt{s}') \sum_s \mu(\mathtt{s})p(\mathtt{s}'|\mathtt{s}) \\
&= \sum_{\mathtt{s}'} \mu(\mathtt{s}')V^2(\mathtt{s}') = \|V\|_D
\end{aligned}$$

□

**Lemma 11.3.** *Let* $\mathtt{w}_{TD}$ *be the limit of* $\mathtt{w}_{\mathtt{t}+1} = \Pi(T^\pi(X\mathtt{w}_\mathtt{t}))$*, i.e., using* $TD(0)$*. Then,*

$$\|X\mathtt{w}_{TD} - \mathcal{V}^\pi\|_D \le \frac{1}{1-\gamma}\|X\mathtt{w}_{min} - \mathcal{V}^\pi\|_D$$

*Proof.* By definition, $\mathtt{w}_{min}$ is the projection of $\mathcal{V}^\pi$ to $X$, i.e., $\Pi(\mathcal{V}^\pi)$. Note that we have $\mathcal{V}^\pi = \Pi(\mathcal{V}^\pi) + (\mathcal{V}^\pi - \Pi(\mathcal{V}^\pi))$ and $\langle \Pi(\mathcal{V}^\pi), \mathcal{V}^\pi - \Pi(\mathcal{V}^\pi)\rangle_D = 0$, otherwise we can improve. This implies that $\|\mathcal{V}^\pi\|_D = \|\Pi(\mathcal{V}^\pi)\|_D + \|\mathcal{V}^\pi - \Pi(\mathcal{V}^\pi)\|_D$.

$$\begin{aligned}
\|X\mathtt{w}_{TD} - \mathcal{V}^\pi\| &\le \|X\mathtt{w}_{TD} - \Pi(\mathcal{V}^\pi\|_D + \|\Pi(\mathcal{V}^\pi) - \mathcal{V}^\pi\|_D \\
&= \|\Pi(T^\pi(X\mathtt{w}_{TD})) - \Pi(\mathcal{V}^\pi)\|_D + \|\Pi(\mathcal{V}^\pi) - \mathcal{V}^\pi\|_D \\
&\le \|T^\pi(X\mathtt{w}_{TD}) - \mathcal{V}^\pi\|_D + \|\Pi(\mathcal{V}^\pi) - \mathcal{V}^\pi\|_D \\
&\le \gamma\|X\mathtt{w}_{TD} - \mathcal{V}^\pi\|_D + \|\Pi(\mathcal{V}^\pi) - \mathcal{V}^\pi\|_D
\end{aligned}$$

and the lemma follows since $\Pi(\mathcal{V}^\pi) = X\mathtt{w}_{min}$ by definition of $\mathtt{w}_{min}$. □

### 11.6.2 LSTD

We consider least Squares Temporal Difference (LSTD). We will analyze $LSTD(0)$.

Let $x_\mathtt{t} = x_{\mathtt{s}_\mathtt{t}}$. The weights:

$$\begin{aligned}
\mathtt{w}_{\mathtt{t}+1} &= \mathtt{w}_\mathtt{t} + \alpha(\mathtt{r}_\mathtt{t} + \gamma\mathtt{w}_\mathtt{t}^\top x_{\mathtt{t}+1} - \mathtt{w}_\mathtt{t}^\top x_\mathtt{t})x_\mathtt{t} \\
\mathtt{w}_{\mathtt{t}+1} &= \mathtt{w}_\mathtt{t} + \alpha(\mathtt{r}_\mathtt{t} x_\mathtt{t} - x_\mathtt{t}(x_\mathtt{t} - \gamma x_{\mathtt{t}+1})^\top \mathtt{w}_\mathtt{t})
\end{aligned}$$

We have that

$$E[\mathtt{w_{t+1}}|\mathtt{w_t}] = \mathtt{w_t} + \alpha(b - A\mathtt{w_t})$$

where $b = E[\mathtt{r_t} x_\mathtt{t}] \in \mathbb{R}^d$ and $A = E[x_\mathtt{t}(x_\mathtt{t} - \gamma x_\mathtt{t+1})^\top] \in \mathbb{R}^{d\times d}$.

At convergence we have $\mathtt{w} = A^{-1}b$.

## 11.7 Applications: games

### 11.7.1 Backgammon

Tesauro developed the TD-gammon in the late 1980's and 1990's. The system uses a neural network approximation, using a $TD(\lambda)$ updates. The initial system was able to win against other computer programs, which where hand-designed, and later matched the performance of the best human. For the training the system used self-play, namely playing against itself.

The neural network has a single hidden layer (with 40/80/160 nodes, in the three different generations of the project). There are 198 inputs. The hidden layer has sigmoidal nodes with $\sigma(z) = \frac{1}{1+e^{-z}}$. The output $y$ can be viewed as the probability of winning from a given position.

The 198 inputs are composed as follows. For each of the 24 slots and for each of the two colors, we have four Boolean inputs. The first indicates if there is a single chip, the second indicates two or more chips, the third indicates exactly three chips and the fourth indicates four or more chips. We have two Boolean variables that indicate who turn it is (one for white and one for black). Finally, we have two aggregate inputs, per color: (1) number of chips divided by 2 and (2) number of chips removed divided by 15.

The TD-gammon uses the standard TD updates.

$$\Delta \mathtt{w_t} = \alpha(y_\mathtt{t} - y_\mathtt{t-1}) \sum_{k=1}^{t} \lambda^{\mathtt{t}-k} \nabla_\mathtt{w} y_k$$

When the game ends we replace $y_\mathtt{t}$ by $z$ the outcome of the game.

The evaluation of the position is done using a max-min tree, whose depth is 1/2/3. The max-min tree has max layer where the value of a node is the maximum of their children, and min layer, where the value of a node is the minimum of their children. Given the tree (of size at most 20/400/8000 in different generations) we evaluate the leaves using the neural network, and propagate the values in the min-max tree to get the predicted value of a position.

The network is initialized with small random weights. Note that due to the game nature, even with random policies the game terminates. The training is done using self-play, where the number of games was $300K/800K/1500K$ in different generations.

One very intriguing aspect of the TD-gammon is that it *improved the human performance*. For example, when the first role is `4-1`, grandmasters did the 1 move in location 6, while the TD-gammon moved location 24. Following the TD-gammon, grandmasters changed their view and started playing those positions in the same way that TD-gammon did.

### 11.7.2 Attari games

The system was developed by DeepMind researchers during $2013-2015$. The system learns to play from the raw video frames, a variety of 49 games. The learning uses a neural network and employs $Q$-learning (and a Deep-Q-Network, DQN).

The neural network works as follows. There is a preprocessing stage that maps frames of size $210 \times 160$ with 128-colors, to $84 \times 84$ and 4-colors. This mapping is fixed and shared by all games. The neural network has three convolutional layers followed by one completely connected layer. The output layer has multiple outputs. The nodes in the network use RelU gates.

The DQN innovations are: (1) Experience reply, reusing the same example many times, (2) fixed Q-target, when setting the target update, and (3) clipping the error. To better understand the innovations, we need to consider the flow of the training.

In the training, we first select an action $\mathtt{a_t}$, using $\varepsilon$-greedy policy. We observe the reward and the next state $(\mathtt{s_t}, \mathtt{a_t}, \mathtt{r_t}, \mathtt{s_{t+1}})$, and store them in the reply memory. We sample a mini-batch from the reply memory. For each sample we use the fixed-Q-target (more latter). We use the SGD to minimize the MSE.

For the fixed-Q-target, we do the update as follows. We have two sets of weights $\mathtt{w_t}$, the current weights, and $\mathtt{w_t^-}$ as the fixed weights. We update $\mathtt{w_t}$ to $\mathtt{w_t^-}$ every (large) number of plays, setting $\mathtt{w_t} = \mathtt{w_t^-}$. The update at time $\mathtt{t}$ becomes:

$$\Delta \mathtt{w_t} = \alpha \Gamma_{\mathtt{t}} \nabla_{\mathtt{w_t}} Q(\mathtt{s_t}, \mathtt{a_t}; \mathtt{w_t})$$

where

$$\Gamma = \mathtt{r_t} + \gamma \max_{\mathtt{a}} Q(\mathtt{s_{t+1}}, \mathtt{a}; \mathtt{w_t^-}) - Q(\mathtt{s_t}, \mathtt{a_t}; \mathtt{w_t})$$

Note that $\mathtt{w_t^-}$ is used only to estimate the future return from $\mathtt{s_{t+1}}$. The clipping enforces $\Gamma_{\mathtt{t}} \in [-1, +1]$.

The system is able to match human performance in 29 out of 49 Attari games.

## 11.8 Bibliography Remarks

The work of Gerald Tesauro on TD-gammon is described in [15, 16].

The DeepMind system for playing Attari games is described in [9].

Part of the outline borrows from David Silver class notes and the the book of Sutton and Barto [14].

# Chapter 12

# Large state space: Policy Gradient Methods

This chapter continues looking at the case where the MDP models are large state space. In the previous chapter we looked at approximating the value function. In this chapter we will consider learning directly a policy and optimizing it.

The main advantages and disadvantages of policy optimization (rather than value function approximation) are the following:

1. Continuous action space. In such a case the policy optimization method is the most natural and the one which is used in practice (for applications such as robotics).

2. Convergence: While the policy optimization would normally converge, it will most likely converge to a local optimum.

3. High dimensional spaces: It can be fairly effective in selecting the actions (in contrast to learning accurate values).

4. Evaluation time would typically be longer (compared to value function approaches).

5. Stochastic Policy: allows naturally to have a stochastic policy.

## 12.1 Stochastic policies

We have seen that the optimal policy in an MDP is deterministic, so what benefit can be in considering a stochastic policy?

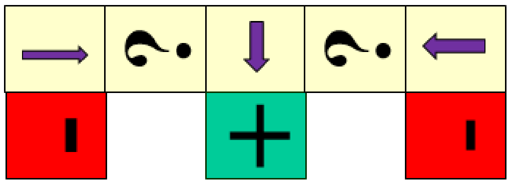


Figure 12.1: Grid-world example

The main benefit is in situations where there is a miss-specification of the model. The main issue is that the state encoding might create a system which is not Markovian anymore, by coalescing certain states which have identical encoding. We will give two example of this phenomena.

### Aliased Grid-world

Consider the example in Figure 12.1. The green state is the good goal and the red ones are the bad. The encoding of each state is the location of the walls. In each state we need to choose a direction. The problem is that we have two states which are indistinguishable (marked by question mark).

It is not hard to see that any deterministic policy would fail from some start state (either the left or the right one). Alternatively, we can use a randomized policy in those states,with probability half go right and probability half go left. For such a policy we have a rather short time to reach the green goal state (and avoid the red states).

The issue here was that two different states had the same encoding, and thus violated the Markovian assumption. This can occur when we encode the state with a small set of features, and some (hopefully, similar) states coallesce to a single representation.

### Zero-sum games

The MDP model was not design for interactive zero-sum games, however, in many of the applications we saw, we train a policy to play a board game (such as backgammon). Any board game is an instance of a zero-sum game, since if one player wins the other loses. In alternating-moves games, the optimal policy would still be deterministic, but this is not the case in simultaneous-move games.

Consider a penny-matching game, in which each player simultaneously select a bit $\{0, 1\}$. If the two selected bits are identical the first player wins and if they differ the second player wins. The best policy for each player is stochastic (selecting each bit with probability half).

An important observation is that if one of the player plays deterministically (or almost deterministically), then the other player can win (or almost always win) by selecting (deterministically) the appropriate action. For this reason, even any $\varepsilon$-greedy would have a poor performance.

Here we violated the assumption that the rewards depend only on the state. In this example they depend indirectly on the policy selected.

## 12.2 Policy optimization

We will assume that each state $\mathtt{s}$ has an encoding $\phi(\mathtt{s}) \in \mathbb{R}^{d_1}$. The policy will have a parametrization $\theta \in \mathbb{R}^{d_2}$. The policy will be based on the two encodings, and we have $\pi(\mathtt{a}|\mathtt{s}, \theta)$, which is the probability of selecting action $\mathtt{a}$ when observing state $\mathtt{s}$ (or, equivalently, $\phi(\mathtt{s})$), and having a policy parametrization $\theta$.

The optimization problem would be:

$$\theta^* = \arg\max_\theta J(\theta)$$

where $J(\theta) \triangleq \mathcal{V}^\pi(\mathtt{s}_0)$ is the expected return of the policy $\pi(\cdot|\cdot, \theta)$ from the initial state $\mathtt{s}_0$. This maximization problem can be solved in multiple ways. We will mainly explore gradient based methods.

We start by giving a few examples on how to parameterize the policy. The first is a *log linear policy*. We will assume an encoding of the state and action pairs, i.e., $\phi(\mathtt{s}, \mathtt{a})$. The linear part will compute $\mu(\mathtt{s}, \mathtt{a}) = \phi(\mathtt{s}, \mathtt{a})^\top \theta$. Given the values of $\mu(\mathtt{s}, \mathtt{a})$ for each $\mathtt{a} \in \mathcal{A}$, the policy select action $\mathtt{a}$ with probability proportional to $e^{\mu(\mathtt{s},\mathtt{a})}$. Namely,

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

The second example is a *Gaussian policy*, where the action space is a real number, i.e., $A = \mathbb{R}$. The encoding is of states, and the actions are any real number. Given a state $\mathtt{s}$ we compute $\mu(\mathtt{s}) = \phi(\mathtt{s})^\top \theta$. We select an action $\mathtt{a}$ from the normal distribution with mean $\mu(\mathtt{s})$ and variance $\sigma^2$, i.e., $N(\mu(\mathtt{s}), \sigma^2)$. (The Gaussian policy has an additional parameter $\sigma$.)

We would like to use the policy gradient to optimize the expected return $J(\theta)$ of the policy $\pi(\cdot|\cdot, \theta)$. We will compute the gradient of $J(\theta)$, i.e., $\nabla_\theta J(\theta)$. The update of the policy parameter $\theta$ is

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

where $\alpha$ is a learning rate.

One challenge we will have to address is to relate $\nabla_\theta J(\theta)$ to the gradients $\nabla\pi(\mathtt{a}|\mathtt{s},\theta)$.

### 12.2.1 Finite differences methods

This methods can be used even when we do not have a representation of the gradient of the policy or even the policy itself. This may arise many times when we have, for example, access to an off-the-shelf robot for which the software is encoded already in the robot. In such cases we can estimate the gradient by introducing perturbations in the parameters.

The simplest case is component-wise gradient estimates. Let $e_i$ be a unit vector, i.e., has in the $i$-th entry a value 1 and a value 0 in all the other entries. The perturbation that we will add is $\delta e_i$ for some $\delta > 0$. We will use the following approximation:

$$\frac{\partial}{\partial\theta_i}J(\theta) \approx \frac{\hat{J}(\theta+\delta e_i) - \hat{J}(\theta)}{\delta}$$

where $\hat{J}(\theta)$ is unbiased estimator of $J(\theta)$. A more symmetric approximation is sometimes better,

$$\frac{\partial}{\partial\theta_i}J(\theta) \approx \frac{\hat{J}(\theta+\delta e_i) - \hat{J}(\theta-\delta e_i)}{2\delta}$$

The problem is that we need to average many samples of $\hat{J}(\theta \pm \delta e_i)$ to overcome the noise. Another weakness is that we need to do the computation per dimension. In addition, the selection of $\delta$ is also critical. A small $\delta$ might have a large noise rate that we need to overcome (by using many samples). A large $\delta$ run the risk of facing the non-linearity of $J$.

Rather then doing the computation per dimension, we can do a more global approach and use a least squares estimation of the gradient. Consider a random vector $u_i$, then we have

$$J(\theta+\delta u_i) \approx J(\theta) + \delta(u_i)^\top \nabla J(\theta)$$

We can define the following least square problem,

$$G = \arg\min_x \sum_i (J(\theta+\delta u_i) - J(\theta) - \delta(u_i)^\top x)^2,$$

where $G$ is our estimate for $\nabla J(\theta)$.

We can move to matrix notation and define $\Delta J^{(i)} = J(\theta+\delta u_i) - J(\theta)$ and $\Delta J = [\cdots, \Delta J^{(i)}, \cdots]^\top$. We define $\Delta\theta^{(i)} = \delta u_i$, and the matrix $[\Delta\Theta] = [\cdots \Delta\theta^{(i)}, \cdots]^\top$, where the $i$-th row is $\Delta\theta^{(i)}$.

We would like to solve for the gradient, i.e,

$$\Delta J \approx [\Delta\Theta] x$$

This is a standard least square problem and the solution is

$$G = ([\Delta\Theta]^\top [\Delta\Theta])^{-1} [\Delta\Theta]^\top \Delta J$$

One issue that we neglected is that we actually do not a have the value of $J(\theta)$. The solution is to solve also for the value of $J(\theta)$. We can define a matrix $M = [1, [\Delta\Theta]]$, a vector of unknowns $x = [J(\theta), \nabla J(\theta)]$, and have the target be $z = [\cdots, J(\theta+\delta u_i), \cdots]$. We can now solve for $z \approx Mx$, and this will recover an estimate also for $J(\theta)$.

## 12.3 Policy Gradient Theorem

The policy gradient theorem will relate the gradient of the expected return $\nabla J(\theta)$ and the gradients of the policy $\nabla\pi(\mathtt{a}|\mathtt{s},\theta)$. We will mainly try to make sure that we are able to use it to get estimates, and the quantities would be indeed observable by the learner.

We will state the policy gradient theorem for the finite-horizon return, but it also holds for the discounted return and average reward return. Recall that the finite-horizon return is $\sum_{\mathtt{t}=1}^{\mathtt{T}} \mathtt{r_t}$ and $\mathcal{V}^\pi(\mathtt{s}) = E[\sum_{\mathtt{t}=1}^{\mathtt{T}} \mathtt{r_t} | \mathtt{s}_1 = \mathtt{s}]$.

**Theorem 12.1** (Policy Gradient Theorem). *For any policy $\pi(\cdot|\cdot;\theta)$, we have*

$$\nabla J(\theta) \propto \sum_{s\in S} \mu(\mathtt{s}) \sum_{\mathtt{a}\in\mathcal{A}} Q^\pi(\mathtt{s},\mathtt{a}) \nabla_\theta \pi(\mathtt{a}|\mathtt{s};\theta)$$

*where $\mu(\mathtt{s}) = \nu(\mathtt{s})/\mathtt{T}$ and $\nu(\mathtt{s}) = \sum_{\mathtt{t}=1}^{\mathtt{T}} \Pr[\mathtt{s_t} = \mathtt{s}|\mathtt{s}_1, \theta]$.*

*Proof.* For each state $\mathtt{s}$ we have

$$\begin{aligned}
\nabla\mathcal{V}^\pi(\mathtt{s}) &= \nabla\sum_{\mathtt{a}}\pi(\mathtt{a}|\mathtt{s})Q^\pi(\mathtt{s},\mathtt{a})\\
&= \sum_{\mathtt{a}} Q^\pi(\mathtt{s},\mathtt{a})\nabla\pi(\mathtt{a}|\mathtt{s}) + \pi(\mathtt{a}|\mathtt{s})\nabla Q^\pi(\mathtt{s},\mathtt{a})\\
&= \sum_{\mathtt{a}} Q^\pi(\mathtt{s},\mathtt{a})\nabla\pi(\mathtt{a}|\mathtt{s}) + \pi(\mathtt{a}|\mathtt{s})\sum_{\mathtt{s}_1} p(\mathtt{s}_1|\mathtt{s},\mathtt{a})\nabla\mathcal{V}^\pi(\mathtt{s}_1)\\
&= \sum_{\mathtt{a}} Q^\pi(\mathtt{s},\mathtt{a})\nabla\pi(\mathtt{a}|\mathtt{s}) + \sum_{\mathtt{s}_1} p(\mathtt{s}_1|\mathtt{s})\nabla\mathcal{V}^\pi(\mathtt{s}_1)\\
&= \sum_{\mathtt{a}} Q^\pi(\mathtt{s},\mathtt{a})\nabla\pi(\mathtt{a}|\mathtt{s}) + \sum_{\mathtt{s}_1} p(\mathtt{s}_1|\mathtt{s})\sum_{\mathtt{a}} Q^\pi(\mathtt{s}_1,\mathtt{a})\nabla\pi(\mathtt{a}|\mathtt{s}_1) + \sum_{\mathtt{s}_2} p(\mathtt{s}_2|\mathtt{s}_1)p(\mathtt{s}_1|\mathtt{s})\nabla\mathcal{V}^\pi(\mathtt{s}_2)\\
&= \sum_{x\in S}\sum_{\mathtt{t}=1}^{\mathtt{T}} \Pr[\mathtt{s}_\mathtt{t}=x|\mathtt{s}_1=\mathtt{s},\pi]\sum_{\mathtt{a}} Q^\pi(x,\mathtt{a})\nabla\pi(\mathtt{a}|x)
\end{aligned}$$

where the first identity follows since by averaging $Q^\pi(\mathtt{s},\mathtt{a})$ over the actions $\mathtt{a}$ with the probabilities induce by $\pi(\mathtt{a}|\mathtt{s})$ we have right expectation of the immediate reward. The next state is distributed correctly, and therefore the identity holds. The second equality follows from the gradient of a multiplication. The third follows since $\nabla Q^\pi(\mathtt{s},\mathtt{a}) = \nabla[\mathtt{r}(\mathtt{s},\mathtt{a}) + \gamma\sum_{\mathtt{s}'} p(\mathtt{s}'|\mathtt{s},\mathtt{a})\mathcal{V}^\pi(\mathtt{s}'|\mathtt{s},\mathtt{a})]$. The next two identities role the policy one step in to the future. The last identity follows from unrolling $\mathtt{s}'$ to $\mathtt{s}''$, etc. and then reorganizing the terms.

Using this we have

$$\begin{aligned}
\nabla J(\theta) &= \nabla\mathcal{V}^\pi(\mathtt{s}_0)\\
&= \sum_{\mathtt{s}}\left(\sum_{\mathtt{t}=1}^{\mathtt{T}}\Pr[\mathtt{s}_\mathtt{t}=\mathtt{s}|\mathtt{s}_0,\pi]\right)\sum_{\mathtt{a}} Q^\pi(\mathtt{s},\mathtt{a})\nabla\pi(\mathtt{a}|\mathtt{s})\\
&= \sum_{\mathtt{s}}\nu(\mathtt{s})\sum_{\mathtt{a}} Q^\pi(\mathtt{s},\mathtt{a})\nabla\pi(\mathtt{a}|\mathtt{s})\\
&= \mathtt{T}\sum_{s}\mu(\mathtt{s})\sum_{\mathtt{a}} Q^\pi(\mathtt{s},\mathtt{a})\nabla\pi(\mathtt{a}|\mathtt{s})
\end{aligned}$$

□

**Example 12.1.** *Consider an MDP with a single state* $\mathtt{s}$ *(which is also called Multi-Arm Bandit, and is discussed in Chapter 15). Assume we have only two actions, action* $\mathtt{a}_1$ *has expected reward* $\mathtt{r}_1$ *and action* $\mathtt{a}_2$ *has expected reward* $\mathtt{r}_2$*.*

*The policy $\pi$ is define with a parameter $\theta = (\theta_1, \theta_2)$, where $\theta_i \in \mathbb{R}$. Given $\theta$ the probability of action $\mathtt{a}_i$ is $p_i = e^{\theta_i}/(e^{\theta_1} + e^{\theta_2})$. We will also select a horizon of length one, i..e, $\mathtt{T} = 1$. This implies that $Q^\pi(\mathtt{s}, \mathtt{a}_i) = \mathtt{r}_i$.*

*In this simple case we can compute directly $J(\theta)$ and $\nabla J(\theta)$. The expected return is simply,*

$$J(\theta) = p_1\mathtt{r}_1 + p_2\mathtt{r}_2 = \frac{e^{\theta_1}}{e^{\theta_1} + e^{\theta_2}}\mathtt{r}_1 + \frac{e^{\theta_2}}{e^{\theta_1} + e^{\theta_2}}\mathtt{r}_2$$

*Note that $\frac{\partial}{\partial\theta_1}p_1 = p_1 - p_1^2 = p_1(1-p_1)$ and $\frac{\partial}{\partial\theta_2}p_1 = -p_1p_2 = -p_1(1-p_1)$. The gradient is*

$$\nabla J(\theta) = \mathtt{r}_1 \begin{pmatrix} p_1(1-p_1) \\ -p_1(1-p_1) \end{pmatrix} + \mathtt{r}_2 \begin{pmatrix} -p_1(1-p_1) \\ p_1(1-p_1) \end{pmatrix} = (\mathtt{r}_1 - \mathtt{r}_2)p_1(1-p_1) \begin{pmatrix} +1 \\ -1 \end{pmatrix}$$

*Updating in the direction of the gradient, in the case that $\mathtt{r}_1 > \mathtt{r}_2$, would increase $\theta_1$ and decrease $\theta_2$, which will imply that $p_1$ will converge to 1.*

*To apply the Policy gradient theorem we need to compute the gradient,*

$$\nabla_\theta \pi(\mathtt{a}_1|\mathtt{s};\theta) = \nabla \begin{pmatrix} p_1 \\ 1-p_1 \end{pmatrix} = \begin{pmatrix} p_1(1-p_1) \\ -p_1(1-p_1) \end{pmatrix}$$

*and the policy gradient theorem gives us the same expression,*

$$\nabla J(\theta) \propto \mathtt{r}_1 \nabla\pi(\mathtt{a}_1|\theta) + \mathtt{r}_2\nabla\pi(\mathtt{a}_2|\theta) = \mathtt{r}_1 \begin{pmatrix} p_1(1-p_1) \\ -p_1(1-p_1) \end{pmatrix} + \mathtt{r}_2 \begin{pmatrix} -p_1(1-p_1) \\ p_1(1-p_1) \end{pmatrix}$$

*where we used the fact that there is only a single state $\mathtt{s}$, and that $Q^\pi(\mathtt{s}, \mathtt{a}_i) = \mathtt{r}_i$.*

The Policy Gradient Theorem gives us a way to compute the gradient. We can sample states from the distribution $\mu(\mathtt{s})$ using the policy $\pi$. We still need to resolve the sampling of the action. We are going to observe the outcome of only one action in state $\mathtt{s}$, and the theorem requires summing over all of them! In the following we will slightly modify the theorem so that we will be able to use only the action $\mathtt{a}$ selected by the policy $\pi$, rather than summing over all actions.

Consider the following simple identity,

$$\nabla f(x) = f(x)\frac{\nabla f(x)}{f(x)} = f(x)\nabla \log f(x)$$

This implies that we can restate the Policy Gradient Theorem as the following corollary,

**Corollary 12.2** (Policy Gradient Corollary)**.** *For any policy $\pi(\cdot|\cdot;\theta)$, we have*

$$\nabla J(\theta) \propto \sum_{s\in S}\mu(\mathtt{s})\sum_{\mathtt{a}\in\mathcal{A}}\pi(\mathtt{a}|\mathtt{s};\theta)Q^\pi(\mathtt{s},\mathtt{a})\nabla_\theta \log\pi(\mathtt{a}|\mathtt{s};\theta) = E^\pi[Q^\pi(\mathtt{s},\mathtt{a})\nabla_\theta\log\pi(\mathtt{a}|\mathtt{s};\theta)]$$

*where $\mu(\mathtt{s}) = \nu(\mathtt{s})/\mathtt{T}$ and $\nu(\mathtt{s}) = \sum_{\mathtt{t}=1}^{\mathtt{T}}\Pr[\mathtt{s}_\mathtt{t} = \mathtt{s}|\mathtt{s}_1,\theta]$.*

Note that in the above corollary both the state $\mathtt{s}$ and action $\mathtt{a}$ are sampled using the policy $\pi$. This avoids the need to sum over all actions, and leaves on the action selected by the policy.

We can now apply does the policy gradient to some simple policy class. For the log-linear policy class we have $\log\pi(\mathtt{a}|\mathtt{s}) \propto \phi(\mathtt{s},\mathtt{a})^\top\theta$, and then

$$\nabla\log\pi(\mathtt{a}|\mathtt{s}) = \phi(\mathtt{s},\mathtt{a}) - \sum_b \pi(b|\mathtt{s};\theta)\phi(\mathtt{s},b)$$

and the update is $\Delta\theta \propto \alpha U(\phi(\mathtt{s},\mathtt{a})-v)$ where $E[U] = Q^\pi(\mathtt{s},\mathtt{a})$ and $v = \sum_b \pi(b|\mathtt{s};\theta)\phi(\mathtt{s},b)$.[1]

For Gaussian policy class we have $\nabla\log\pi(\mathtt{a}|\mathtt{s};\theta) \propto (\mathtt{a}-\mu(\mathtt{s}))\phi(\mathtt{s})/\sigma^2$ and update is $\Delta\theta \propto \alpha U(\mathtt{a}-\mu(\mathtt{s}))\phi(\mathtt{s})/\sigma^2$ where $E[U] = Q^\pi(\mathtt{s},\mathtt{a})$.

**Example 12.2.** *Consider the following deterministic MDP. We have states $\mathcal{S} = \{\mathtt{s}_0,\mathtt{s}_1,\mathtt{s}_2,\mathtt{s}_3\}$ and actions $\mathcal{A} = \{\mathtt{a}_0,\mathtt{a}_1\}$. We start at $\mathtt{s}_0$. Action $\mathtt{a}_0$ from any state leads to $\mathtt{s}_3$. Action $\mathtt{a}_1$ moves from $\mathtt{s}_0$ to $\mathtt{s}_1$ and from $\mathtt{s}_1$ to $\mathtt{s}_2$. All the rewards are zero except the terminal reward at $\mathtt{s}_2$ which is 1. The horizon is $\mathtt{T} = 2$. This implies that the optimal policy performs in each state $\mathtt{a}_1$ and has a return of 1.*

*We have a log-linear policy parameterized by $\theta\in\mathbb{R}^4$. In state $\mathtt{s}_0$ it selects action $\mathtt{a}_1$ with probability $p_1 = e^{\theta_1}/(e^{\theta_1}+e^{\theta_2})$, and in state $\mathtt{s}_1$ it selects action $\mathtt{a}_1$ with probability $p_2 = e^{\theta_3}/(e^{\theta_3}+e^{\theta_4})$.*

*For this simple MDP we can specify the expected return $J(\theta) = p_1p_2$. We can also compute the gradient and have*

$$\nabla J(\theta) = \begin{pmatrix} p_1(1-p_1)p_2 \\ -p_1(1-p_1)p_2 \\ p_1p_2(1-p_2) \\ -p_1p_2(1-p_2)\end{pmatrix} = p_1p_2\begin{pmatrix}(1-p_1)\\ -(1-p_1)\\ (1-p_2)\\ -(1-p_2)\end{pmatrix}$$

*The policy gradient theorem will use the following ingredients. The $Q^\pi$ is: $Q^\pi(\mathtt{s}_0,\mathtt{a}_1) = p_2$, $Q^\pi(\mathtt{s}_1,\mathtt{a}_1) = 1$ and all the other entries are zero. The weights of the states are*

[1]YM: Can we really ignore the normalizing term?

*$\mu(\mathtt{s}_0) = 1$, $\mu(\mathtt{s}_1) = p_1$, $\mu(\mathtt{s}_2) = p_1p_2$ and $\mu(\mathtt{s}_3) = 2 - p_1 - p_1p_2$. The gradient of the action in each state is:*

$$\nabla\pi(\mathtt{a}_1|\mathtt{s}_0;\theta) = \begin{pmatrix}1\\0\\0\\0\end{pmatrix} - p_1\begin{pmatrix}1\\0\\0\\0\end{pmatrix} - (1-p_1)\begin{pmatrix}0\\1\\0\\0\end{pmatrix} = (1-p_1)\begin{pmatrix}1\\-1\\0\\0\end{pmatrix}$$

*Similarly*

$$\nabla\pi(\mathtt{a}_1|\mathtt{s}_1;\theta) = \begin{pmatrix}0\\0\\1\\0\end{pmatrix} - p_2\begin{pmatrix}0\\0\\1\\0\end{pmatrix} - (1-p_2)\begin{pmatrix}0\\0\\0\\1\end{pmatrix} = (1-p_2)\begin{pmatrix}0\\0\\1\\-1\end{pmatrix}$$

*The policy gradient theorem states that the expected retrun gradient is*

$$\mu(\mathtt{s}_0)Q^\pi(\mathtt{s}_0,\mathtt{a}_1)\pi(\mathtt{a}_1|\mathtt{s}_0;\theta)\nabla\pi(\mathtt{a}_1|\mathtt{s}_0;\theta) + \mu(\mathtt{s}_1)Q^\pi(\mathtt{s}_1,\mathtt{a}_1)\pi(\mathtt{a}_1|\mathtt{s}_1;\theta)\nabla\pi(\mathtt{a}_1|\mathtt{s}_1;\theta)$$

*where we dropped all the terms that evaluate to zero. plugging in our values we have*

$$p_2p_1(1-p_1)\begin{pmatrix}1\\-1\\0\\0\end{pmatrix} + p_1p_2(1-p_2)\begin{pmatrix}0\\0\\1\\-1\end{pmatrix} = p_1p_2\begin{pmatrix}(1-p_1)\\-(1-p_1)\\(1-p_2)\\-(1-p_2)\end{pmatrix}$$

*which is identical to $\nabla J(\theta)$.*

The main benefit of the policy gradient theorem is that we do not need to compute either $\mu(\mathtt{s})$ or $Q^\pi$. We simply run policy $\pi$ which visits each state $\mathtt{s}$ an expected $\mu(\mathtt{s})$ times, and let $\pi$ select actions. We substitute for $Q^\pi$ and unbiased random variable with the same value. The only thing that we need to explicitly compute in the gradient $\nabla\pi(\mathtt{a}|\mathtt{s};\theta)$ which depends on our parametrization. In the next section we couple the policy gradient theorem with Monte-Carlo updates to derive the REINFORCE algorithm.

## 12.4 REINFORCE: Monte-Carlo updates

The REINFORCE algorithm uses a Monte-Carlo updates in conjunction with the policy gradient computation. Given a episode $(\mathtt{s}_1,\mathtt{a}_1,\mathtt{r}_1,\ldots,\mathtt{s}_T,\mathtt{a}_T,\mathtt{r}_T)$ for each $\mathtt{t}\in[1,T]$ updates,

$$\theta \leftarrow \theta + \alpha R_{\mathtt{t}:T}\nabla\log\pi(\mathtt{a}_\mathtt{t}|\mathtt{s}_\mathtt{t};\theta)$$

where $R_{\mathtt{t}:T} = \sum_{i=\mathtt{t}}^{T}\mathtt{r}_i$. (We are using here every-visit updates, however, sice we have a large state space, it is likely that we never observe the same state twice.)

## baseline function

We can extend the REINFORCE to add a baseline function. The baseline function $b(\mathtt{s})$ can depend in an arbitrary way on the state, but does not depend on the action. The main observation would be that we can add or subtract any such function from our estimate $U$, and it will still be unbiased. This follows since

$$\sum_{\mathtt{a}} b(\mathtt{s})\nabla\pi(\mathtt{a}|\mathtt{s};\theta) = b(\mathtt{s})\nabla\sum_{\mathtt{a}}\pi(\mathtt{a}|\mathtt{s};\theta) = b(\mathtt{s})\nabla 1 = 0$$

Given this, we can restate the Policy Gradient Theorem as,

$$\nabla J(\theta) \propto \sum_{s\in S}\mu(\mathtt{s})\sum_{\mathtt{a}\in\mathcal{A}}(Q^\pi(\mathtt{s},\mathtt{a}) - b(\mathtt{s}))\nabla_\theta\pi(\mathtt{a}|\mathtt{s};\theta)$$

This gives us a degree of freedom to select $b(\mathtt{s})$. Note that by setting $b(\mathtt{s}) = 0$ we get the original theorem. In many cases it is reasonable to use for $b(\mathtt{s})$ the value of the state, i.e., $b(\mathtt{s}) = \mathcal{V}^\pi(\mathtt{s})$. The motivation for this is to reduce the variance of the estimator. If we assume that the magnitude of the gradients $\|\nabla\pi(\mathtt{a}|\mathtt{s};\theta)\|$ is similar for all action $\mathtt{a}\in\mathcal{A}$, we are left with $E^\pi[(Q^\pi(\mathtt{s},\mathtt{a}) - b(\mathtt{s}))^2]$ which is minimized by $b(\mathtt{s}) = E^\pi[Q^\pi(\mathtt{s},\mathtt{a})] = \mathcal{V}^\pi(\mathtt{s})$.

We are left with the challenge of approximating $\mathcal{V}^\pi(\mathtt{s})$. On the one hand this is part of the learning. On the other hand we have developed tools to address this in the previous lecture on value function approximation. We can use $\mathcal{V}^\pi(\mathtt{s}) \approx V(\mathtt{s};\mathtt{w}) = b(\mathtt{s})$. The good news is that any $b(\mathtt{s})$ will keep the estimator unbiased, so we do not depend on $V(\mathtt{s};\mathtt{w})$ to be unbiased.

We can now describe the REINFORCE algorithm with baseline function. We will use a Monte-Carlo sampling to estimate $\mathcal{V}^\pi(\mathtt{s})$ and this will define our function $b(\mathtt{s})$. We will update using $U - b(\mathtt{s})$ where $E[U] = Q^\pi(\mathtt{s},\mathtt{a})$. More specifically, our algorithm will have a class of value approximation function $V(\cdot;\mathtt{w})$ and a parameterized policy $\pi(\cdot|\cdot;\theta)$, and in addition to step size parameters, $\alpha,\beta > 0$. Given an episode $(\mathtt{s}_1,\mathtt{a}_1,\mathtt{r}_1,\ldots,\mathtt{s}_T,\mathtt{a}_T,\mathtt{r}_T)$ for each $\mathtt{t}\in[1,T]$ we compute the $R_{\mathtt{t}:T}$ during the times $[\mathtt{t},T]$, i.e., $R_{\mathtt{t}:T} = \sum_{i=t}^T \mathtt{r}_i$. The error in time $\mathtt{t}$ is $\Gamma_\mathtt{t} = R_{\mathtt{t}:T} - V(\mathtt{s}_\mathtt{t};\mathtt{w})$. The updates are

$$\Delta w = \alpha\Gamma_\mathtt{t}\nabla V(\mathtt{s}_\mathtt{t};\mathtt{w})$$

and

$$\Delta\theta = \beta\Gamma_\mathtt{t}\nabla\log\pi(\mathtt{a}_\mathtt{t}|\mathtt{s}_\mathtt{t};\theta)$$

We can extend this to handle also TD updates. We will use an actor-critic algorithm. We will use a $Q$-value updates for this (but can be done similarly with $V$-values).

The critic maintains an approximate $Q$ function $Q(\mathtt{s},\mathtt{a};\mathtt{w})$. For each time $\mathtt{t}$ it defines the TD error to be $\Gamma_\mathtt{t} = \mathtt{r}_\mathtt{t} + Q(\mathtt{s}_{\mathtt{t+1}},\mathtt{a}_{\mathtt{t+1}};\mathtt{w}) - Q(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t};\mathtt{w})$. The update will be $\Delta w = \alpha\Gamma\nabla Q(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t};\mathtt{w})$. The critic send the actor the TD error $\Gamma$.

The actor maintains a policy $\pi$ which is parameterized by $\theta$. Given a TD error $\Gamma$ it updates $\Delta\theta = \beta\Gamma\nabla\log\pi(\mathtt{a}_\mathtt{t}|\mathtt{s}_\mathtt{t};\theta)$. Then it selects $\mathtt{a}_{\mathtt{t+1}} \sim \pi(\cdot|\mathtt{s}_{\mathtt{t+1}};\theta)$.

We need to be careful in the way we select the function approximation $Q(\cdot;\mathtt{w})$ since it might introduce a bias. The following theorem identifies a special case which guarantee thats we will not have such a bias.

Let the expected square error of $\mathtt{w}$ is

$$SE(\mathtt{w}) = \frac{1}{2}E^\pi[(Q^\pi(\mathtt{s},\mathtt{a}) - Q(\mathtt{s},\mathtt{a};\mathtt{w}))^2]$$

A value function is *compatible* if,

$$\nabla_\mathtt{w} Q(\mathtt{s},\mathtt{a};\mathtt{w}) = \nabla_\theta\pi(\mathtt{a}|\mathtt{s};\theta)$$

**Theorem 12.3.** *Assume that $Q$ is compatible and $\mathtt{w}$ minimizes $SE(\mathtt{w})$, then,*

$$\nabla_\theta J(\theta) \propto E^\pi[Q(\mathtt{s},\mathtt{a};\mathtt{w})\nabla\log\pi(\mathtt{a}|\mathtt{s};\theta)]$$

*Proof.* Since $\mathtt{w}$ minimizes $SE(\mathtt{w})$ we have

$$\begin{aligned} 0 &= \nabla_\mathtt{w} SE(\mathtt{w}) \\ &= \nabla_\mathtt{w} E^\pi[(Q^\pi(\mathtt{s},\mathtt{a}) - Q(\mathtt{s},\mathtt{a};\mathtt{w}))^2] \\ &= E^\pi[(Q^\pi(\mathtt{s},\mathtt{a}) - Q(\mathtt{s},\mathtt{a};\mathtt{w}))\nabla_\mathtt{w} Q(\mathtt{s},\mathtt{a};\mathtt{w})] \end{aligned}$$

Since $Q$ is compatible, we have $\nabla_\mathtt{w} Q(\mathtt{s},\mathtt{a};\mathtt{w}) = \nabla_\theta\pi(\mathtt{a}|\mathtt{s};\theta)$ which implies,

$$0 = E^\pi[(Q^\pi(\mathtt{s},\mathtt{a}) - Q(\mathtt{s},\mathtt{a};\mathtt{w}))\nabla_\theta\log\pi(\mathtt{a}|\mathtt{s};\theta)]$$

and have

$$E^\pi[Q^\pi(\mathtt{s},\mathtt{a})\nabla_\theta\log\pi(\mathtt{a}|\mathtt{s};\theta)] = E^\pi[Q(\mathtt{s},\mathtt{a};\mathtt{w})\nabla_\theta\log\pi(\mathtt{a}|\mathtt{s};\theta)]$$

This implies that by substituting $Q$ in the policy gradient theorem we have

$$\nabla_\theta J(\theta) \propto E^\pi[Q(\mathtt{s},\mathtt{a};\mathtt{w})\nabla\log\pi(\mathtt{a}|\mathtt{s};\theta)]$$

$\square$

We can summarize the various updates for the policy gradient as follows:

- REINFORCE (which is a Monte-Carlo estimate) uses $E^\pi[R_\mathtt{t} \nabla \log \pi(\mathtt{a}|\mathtt{s};\theta)]$.
- Q-function with actor-critic uses $E^\pi[Q(\mathtt{a_t}|\mathtt{s_t};\mathtt{w}) \nabla \log \pi(\mathtt{a}|\mathtt{s};\theta)]$.
- A-function with actor-critic uses $E^\pi[A(\mathtt{a_t}|\mathtt{s_t};\mathtt{w}) \nabla \log \pi(\mathtt{a}|\mathtt{s};\theta)]$, where $A(\mathtt{a}|\mathtt{s};\mathtt{w}) = Q(\mathtt{s},\mathtt{a};\mathtt{w}) - V(\mathtt{s};\mathtt{w})$.
- TD with actor-critic uses $E^\pi[\Gamma \nabla \log \pi(\mathtt{a}|\mathtt{s};\theta)]$, where $\Gamma$ is the TD error.

## 12.5 Application

### 12.5.1 RoboSoccer: training Aibo

Aibo is a small robot that was used in the Robo-Soccer competitions. One of the main tasks was to get the robot to walk fast and stable. For this the UT Austin team used reinforcement learning, and specifically policy gradient.

The robot is controlled through 12 parameters which include: (1) For the front and rear legs tree parameters: height, x-pos., y-pos. (2) For the locus: length/skew multiplier. (3) height of the body both front and rear. (4) Time (per foot) to go through locus, and (5) time (per foot) on ground or in the air.

The training was done in episodes. The Aibo was trained to walk between two landmarks. (See Figure 12.2.) The optimization used a Policy gradient improvement using Finite difference method. Note that since the actual policy is unknown, there are not that many alternatives.

Given a set of parameters $\theta$, there was a perturbation introduced in the following way. The value of $\theta_i$ was modified to $\theta_i + z_i$ where $z_i$ is selected at random from $\{+\varepsilon_i, 0, -\varepsilon_i\}$. The values of the $\varepsilon_i$ we selected to be small compared to the value of $\theta_i$. After the perturbation we have a new vector $\theta'$. Many such vectors are sampled. For each vector the policy is evaluated (the return is the time to walk a certain distance, and we like to minimize the time).

After running many such polices, the update is done as follows. For each attribute $i \in [1, 12]$, we split the policies to three subsets, according to the value of $z_i$, and compute the average return in each subset. If the best outcome is for $z_i = 0$ then we set $A_i = 0$. Otherwise, we set $A_i = avg_i(+\varepsilon_i) - avg_i(-\varepsilon_i)$. The parameter vector is set to

$$\theta \leftarrow \theta + \alpha \frac{A}{\|A\|}$$

(See also Figure 12.3.)

More figures and data are in the slides.

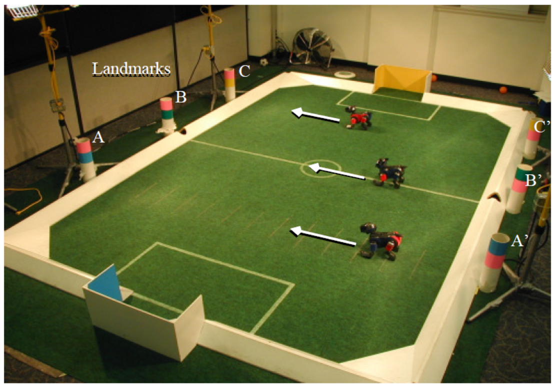
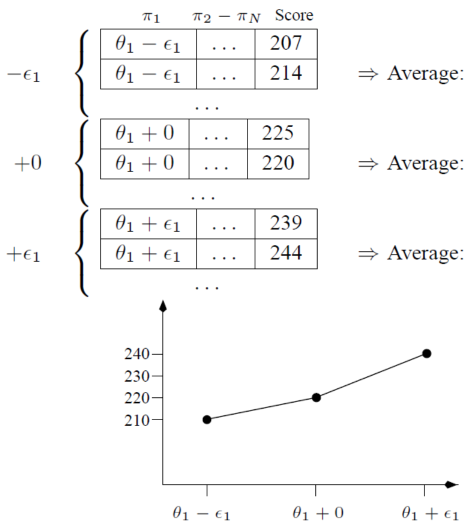


Figure 12.2: The Aibo training environment



| | $\pi_1$ | $\pi_2 - \pi_N$ | Score | |
|---|---|---|---|---|
| $-\epsilon_1$ | $\theta_1 - \epsilon_1$ | ... | 207 | $\Rightarrow$ Average: 210 |
| | $\theta_1 - \epsilon_1$ | ... | 214 | |
| | ... | | | |
| $+0$ | $\theta_1 + 0$ | ... | 225 | $\Rightarrow$ Average: 220 |
| | $\theta_1 + 0$ | ... | 220 | |
| | ... | | | |
| $+\epsilon_1$ | $\theta_1 + \epsilon_1$ | ... | 239 | $\Rightarrow$ Average: 240 |
| | $\theta_1 + \epsilon_1$ | ... | 244 | |
| | ... | | | |

Figure 12.3: The policy updates

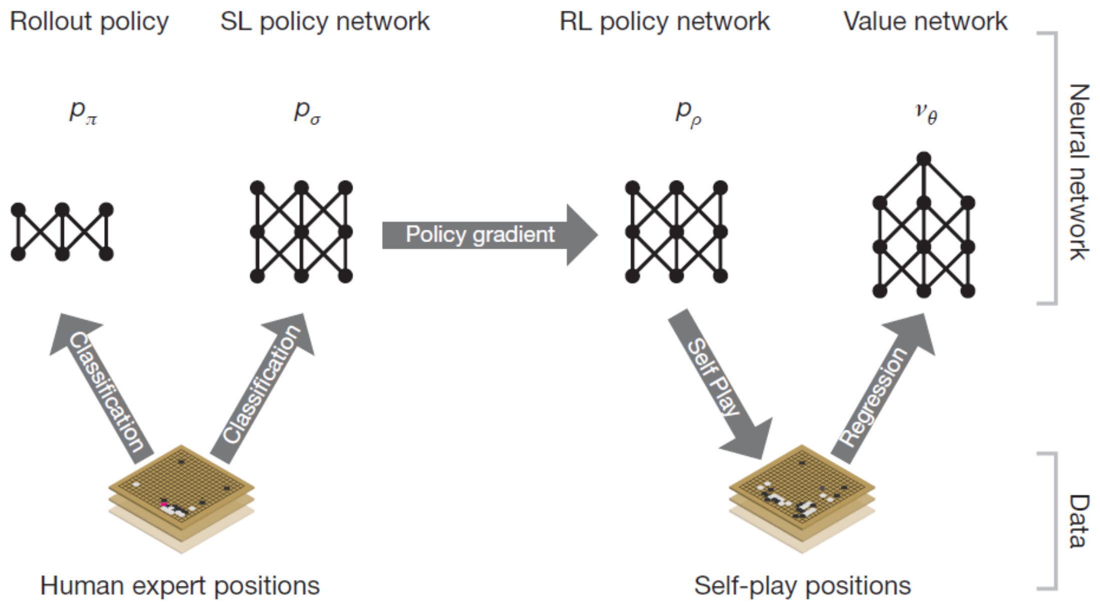


Figure 12.4: The AlphaGo training process

## 12.5.2 AlphaGo

AlphaGo and its successor AlphaGo Zero are the most advanced computer Go players, and the first to beat the best professional players. We will concentrate on AlphaGo. AlphaGo Zero has an additional benefit that it does not have any prior information about the game (mainly in the form of games between human experts).

The training has roughly three phases. The first phase learns from from human games (played by experts). Those games are used to derive a network SL (Supervised Learning) to predict next move of the human, and a Rollout which is a simple but very fast prediction (3 orders faster than SL).

During phase two an RL network is trained. It is initialized with SL weights, and is optimized on self-play against itself.

In phase three, a value network is trained to predict the winner of the game (trained on the RL network self-play games). (See Figure 12.4.)

**Phase 1**

The SL network is trained to predict the human moves. It is a deep network (13 layers), and the goal is to output $p(\mathtt{a}|\mathtt{s};\sigma)$ where $\sigma$ are the parameters of the network.

The parameters are update using a gradient step,

$$\Delta\sigma \propto \frac{\partial \log p(\mathtt{a}|\mathtt{s};\sigma)}{\partial\sigma}$$

Overall, this network increased the prediction accuracy (compared to previous computer programs) from 44% to 55%. We should note that the increase in accuracy translated to a much more significant improvement in the performance.

In addition to SL, another very fast classifier Rollout was build. It uses a linear function approximation and predicts using a soft-max. The accuracy is only 24% but it is much faster ($2\mu s$ compared to $3ms$, a factor of 1500).

**Phase 2**

We learn an RL network that outputs $p(\mathtt{a}|\mathtt{s};\rho)$ where $\rho$ are the parameters of the network. The network structure is identical to that of SL, and the RL is initialized to the weights of SL, i.e., $\sigma$.

The RL is trained using self-play. Rather then playing against the most recent network, the opponent is selected at random between the recent RL configurations. This is done to avoid overfitting. The rewards of the game are given only at the end (win or lose).

The training is done using SGD with policy gradient

$$\Delta\rho \propto \frac{\partial \log p(\mathtt{a}|\mathtt{s};\rho)}{\partial\rho}$$

The learned RL network outperformed SL (wins 80%) of the games. However, SL is better in predicting expert human behavior.

**Phase 3**

We learn a value function $V(\mathtt{s};\theta)$. The goal is to predict the probability that RL will win, and it is trained on the self-play data of RL. It uses a SGD update

$$\Delta\theta \propto \frac{\partial V(\mathtt{s};\theta)}{\partial\theta}$$

When the training was done on complete games, there was a serious problem of over-fitting. (The training error was 0.19 while the test error was 0.37). For this reason, the training is limited to only one position per game. The error rate is about 0.22.

The system uses a Monte-Carlo Search Trees (MCST) to evaluate the positions. The tree is used to perform a lookahead. It uses the Rollout policy to generate

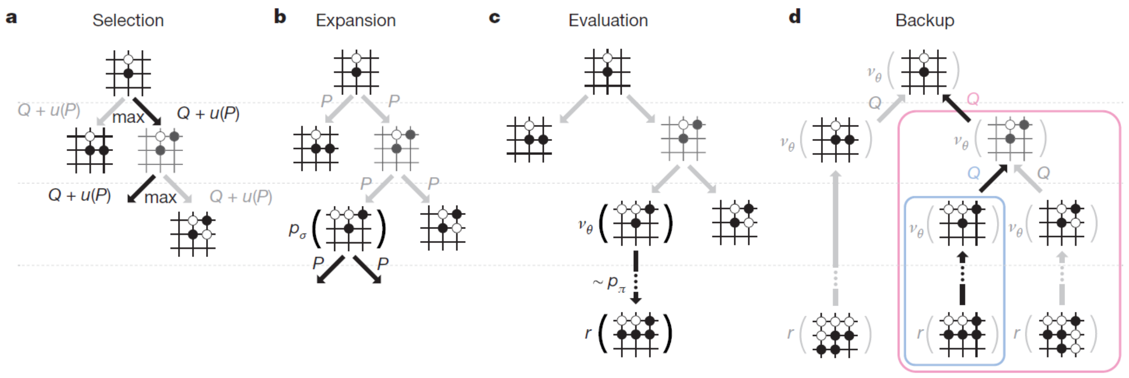


Figure 12.5: AlphaGo use of Monte-Carlo trees

trajectory from the given position. The prediction is an average of the value network prediction $V(\mathbf{s};\theta)$ and the outcome of the Rollout policy. To see a run of the Monte-Carlo Search Tree see Figure 12.5.

## 12.6 Bibliography Remarks

The training of Aibo for RoboSoccer is by [7].

The work on AlphaGo is by [11].

Part of the outline borrows from David Silver class notes and the the book of Sutton and Barto [14].

# Chapter 13

# Large state space: tree based optimization

# Chapter 14

# Deep RL

# Chapter 15

# Multi-Arm bandits

## 15.1 Stochastic Bandits and Regret Minimization

We consider a simplified model of an MDP where there is only a single state and a fixed set $A$ of $k$ actions (a.k.a., arms). We consider a finite horizon problem, where the horizon is $T$. Clearly, the planning problem is trivial, simply select the action with the highest expected reward. We will concentrate on the learning perspective, where the expected reward of each action are unknown.

At each round $1 \leq t \leq T$ the player selects and executes an action. After executing the action, the player observes the reward of the action. However, the rewards of the other actions in $A$ are not revealed to the player.

The reward for action $i$ at round $t$ is denoted by $r_t(i) \sim D_i$, where the support of the reward distribution $D_i$ is $[0, 1]$. We assume that the rewards are i.i.d. (independent and identically distributed).

### Motivation

1. **News**: a user visits a news site and is presented with a news header. The user either clicks on this header or not. The goal of the website is to maximize the number of clicks. So each possible header is an action in a bandit problem, and the clicks are the rewards

2. **Medical Trials**: Each patient in the trial is prescribed one treatment out of several possible treatments. Each treatment is an action, and the reward for each patient is the effectiveness of the prescribed treatment.

3. **Ad selection**: In website advertising, a user visits a webpage, and a learning algorithm selects one of many possible ads to display. If an advertisement is

displayed, the website observes whether the user clicks on the ad, in which case the advertiser pays some amount $v_a \in [0,1]$. So each advertisement is an action, and the paid amount is the reward.

**Model**

- A set of actions $A = \{a_1 \ldots, a_k\}$
- Each action $a_i$ has a reward distribution $D_i$ over $[0,1]$.
- The expectation of distribution $D_i$ is:

$$\mu_i = E_{X \sim D_i}[X]$$

- $\mu^* = \max_i \mu_i$ and $a^* = \arg\max_i \mu_i$.
- $a_t$ is the action the learner chose at round $t$

$$Regret = \max_{i \in A} \sum_{t=1}^{T} \underbrace{r_t(i)}_{\text{Random variable}} - \sum_{t=1}^{T} \underbrace{r_t(a_t)}_{\text{Random variable}}$$

$$\begin{aligned} \text{Pseudo Regret} &= \max_i E\left[\sum_{t=1}^{T} r_t(i)\right] - E\left[\sum_{t=1}^{T} r_t(a_t)\right] \\ &= \mu^* \cdot T - \sum_{t=1}^{T} \mu_{a_t} \end{aligned}$$

We will use extensively the following concentration bound.

**Theorem 15.1** (Hoeffding's inequality)**.** *Given $X_1, \ldots, X_m$ i.i.d random variables s.t $X_i \in [0,1]$ and $E[X_i] = \mu$.*

$$\underbrace{Pr[\frac{1}{m}\sum_{i=1}^{m} X_i - \mu}_{\frac{1}{m}S} \geq \epsilon] \leq \exp(-\tfrac{\epsilon^2 m}{2})$$

### 15.1.1 Warmup: Full information $k = 2$

We start with a simple case where there are two actions and we observe the reward of both actions at each time $t$. We will analyze the greedy policy, which selects the action with the higher average reward (so far).

The greedy policy at time $t$ does the following:

- We observe $\langle r_t(1), r_t(2) \rangle$
- Define

$$avg_t(i) = \frac{1}{t} \sum_{\tau=1}^{t} r_\tau(i)$$

- In time $t+1$ we choose:

$$a_{t+1} = \arg \max_{i \in \{1,2\}} avg_t(i)$$

We now would like to compute the expected regret of the greedy policy. W.l.o.g., we assume that $\mu_1 \geq \mu_2$, and define $\Delta = \mu_1 - \mu_2 \geq 0$.

$$\text{Pseudo Regret} = \sum_{t=1}^{\infty} (\mu_1 - \mu_2) \Pr\left[avg_t(2) \geq avg_t(1)\right]$$

Clearly, at any time $t$,

$$E[avg_t(2) - avg_t(1)] = \mu_2 - \mu_1 = -\Delta$$

We can define a random variable $X_t = r_t(2) - r_t(1) + \Delta$ and $E[X_t] = 0$. Since $(1/t) \sum_t X_t = avg_t(2) - avg_t(1) + \Delta$, by Theorem 15.1

$$Pr[S_2(t) \geq S_1(t)] = Pr\left[avg_t(2) - avg_t(1) + \Delta \geq \Delta\right] \leq e^{-\Delta^2 \frac{t}{2}}$$

We can now bound the regret as follows,

$$\begin{aligned} E\left[\text{Pseudo Regret}\right] &= \sum_{t=1}^{\infty} \Delta \Pr\left[S_2(t) \geq S_1(t)\right] \\ &\leq \sum_{t=1}^{\infty} \Delta e^{-\Delta^2 \frac{t}{2}} \\ &\leq \int_0^{\infty} \Delta e^{-\Delta^2 \frac{t}{2}} dt \\ &= \left[\frac{2}{\Delta} e^{-\Delta^2 \frac{t}{2}}\right]_0^{\infty} \\ &= \frac{2}{\Delta} \end{aligned}$$

Notice that this bound does not depend on $T$!

### 15.1.2 Stochastic Multi-Arm Bandits

We will now see that we cannot get a regret that does not depend on $T$ for the bandits case. Considering the following example:

$$a_1 \sim Br\left(\frac{1}{2}\right)$$

For action $a_2$ there are two alternatives, each with probability $1/2$,

$$a_2 \sim Br\left(\frac{1}{4}\right)\left(w.p. \frac{1}{2}\right) \qquad or \qquad a_2 \sim Br\left(\frac{3}{4}\right)\left(w.p. \frac{1}{2}\right)$$

Assume by way of contradiction

$$E\left[\sum_{i \in \{1,2\}} \Delta_i |T_i|\right] = E\left[PseudoRegret\right] = R$$

where $R$ does not depend on $T$.
By Markov inequality:

$$Pr\left[PseudoRegret \geq 2R\right] \leq \frac{1}{2}$$

Since $\mu_1$ is known, an optimal algorithm will first check $a_2$ in order to decide which action is better and stick with it.

Assuming $\mu_2 = \frac{1}{4}$, and the algorithm decided to stop playing $a_2$ after $M$ rounds, Then:

$$PseudoRegret = \frac{1}{4}M$$

Thus,

$$Pr\left[PseudoRegret \geq 2R\right] = Pr\left[M \geq 8R\right] \leq \frac{1}{2}$$

And,

$$Pr\left[M < 8R\right] > \frac{1}{2}$$

Hence, the probability that after $8R$ rounds, the algorithm will stop playing $a_2$ (if $\mu_2 = \frac{1}{4}$) is at least $\frac{1}{2}$. This implies that there is some sequence of $8R$ outcomes which will result is stopping to try action $a_2$. For simplicity, assume that the sequence is the all zero sequence.

Assume $\mu_2 = \frac{3}{4}$, but all $8R$ first rounds, playing $a_2$ yield the value zero (with probability $\left(\frac{1}{4}\right)^{8R}$). We assumed that after $8R$ zeros for action $a_2$ the algorithm will stop playing $a_2$, even though it is the preferred action. In this case, we will get:

$$PseudoRegret = \frac{1}{4}(T - M) \approx \frac{1}{4}T$$

The expected Pseudo Regret is,

$$E\left[PseudoRegret\right] = R \geq \underbrace{\frac{1}{2}}_{a_2\sim Pr\left(Br\left(\frac{3}{4}\right)\right)} \cdot \underbrace{\left(\frac{1}{4}\right)^{8R}}_{Pr\left(all\ 0|a_2\sim Pr\left(Br\left(\frac{3}{4}\right)\right)\right.} \cdot (T - 8R) \approx e^{-O(R)}T$$

Which implies that:

$$R = O\left(\log T\right)$$

Contrary to the assumption that $R$ does not depend on $T$.

## 15.2 Explore-Then-Exploit

1. We choose a parameter $M$. For $M$ phases we choose each action once (for a total of $kM$ rounds of exploration).

2. After $kM$rounds we always choose the action that had highest average reward during the explore phase.

Define:

$$T_j = \{t : a_t = j, t \leq k \cdot M\}$$
$$\hat{\mu}_j = \frac{1}{M} \sum_{t \in T_j} r_j(t)$$
$$\mu_j = E[r_j(t)]$$
$$\Delta_j = \mu^* - \mu_j$$

where $\Delta_j$ is the difference in expected reward of action $j$ and the optimal action.

We can now write the regret as a function of those parameters:

$$E\left[\text{Pseudo regret}\right] = \underbrace{\sum_{j=1}^{k} \Delta_j \cdot M}_{Explore} + \underbrace{(T - k \cdot M) \sum_{j=1}^{k} \Delta_j Pr\left[j = \arg\max_i \hat{\mu}_i\right]}_{Exploit}$$

For the analysis define:

$$\lambda = \sqrt{\frac{8 \log T}{M}}$$

By Theorem 15.1 we have

$$\Pr\left[|\hat{\mu}_j - \mu_j| \geq \lambda\right] \leq 2e^{-\frac{\lambda M}{2}} = \frac{2}{T^4}$$

which implies (using the union bound) that

$$\Pr \underbrace{\left[\exists_j : |\hat{\mu}_j - \mu_j| \geq \lambda\right]}_{B} \leq \frac{2k}{T^4} \underset{for\, k \leq T}{\leq} \frac{2}{T^3}$$

Define the "bad event" $B = \{\exists_j : |\hat{\mu}_j - \mu_j| \geq \lambda\}$. If $B$ did not happen then for each action $j$, such that $\hat{\mu}_j \geq \hat{\mu}^*$, we have

$$\mu_j + \lambda \geq \hat{\mu}_j \geq \hat{\mu}^* \geq \mu^* - \lambda$$

therefore:

$$2\lambda \geq \mu^* - \mu_j = \Delta_j$$

and therefore:

$$\Delta_j \leq 2\lambda$$

Then, we can bound the expected regret as follows:

$$E[Regret] \leq \underbrace{\left(\sum_{j=1}^{k} \Delta_j\right) M}_{Explore} + \underbrace{(T - k \cdot M) \cdot 2\lambda}_{\text{B didn't happen}} + \underbrace{\frac{2}{T^3} \cdot T}_{\text{B happened}}$$

$$\leq k \cdot M + 2 \cdot \sqrt{\frac{8 \log T}{M}} \cdot T + \frac{2}{T^2}$$

If we optimize the number of exploration phases $M$ and choose $M = T^{\frac{2}{3}}$, we get:

$$k \cdot T^{\frac{2}{3}} + 2 \cdot \sqrt{8 \log T} \cdot T^{\frac{2}{3}} + \frac{2}{T^2}$$

which is sub-linear but more than the $O(\sqrt{T})$ rate we would expect.

## 15.3 Improved Regret Minimization Algorithms

We will look at some more advanced algorithms that mix the exploration and exploitation.

Define:

$n_t(i)$ - the number of times we chose action $i$ by round $t$

$\hat{\mu}_t(i)$ - the average reward of action $i$ so far, that is:

$$\hat{\mu}_t(i) = \sum_{t=1}^{T} r_i(t) \mathbb{I}\left(a_t = i\right) \frac{1}{n_i(t)}$$

Notice that $n_i(t)$ is a random variable and not a number!

We would like to get the following result:

$$\Pr\left[|\hat{\mu}_t(i) - \mu_i| \leq \underbrace{\sqrt{\frac{8 \log T}{n_i(t)}}}_{\lambda_t(i)}\right] \geq 1 - \frac{2}{T^4}$$

We would like to look at the $m^{th}$ time we sampled action $i$:

$$\hat{\mathbb{V}}_m(i) = \frac{1}{m} \sum_{\tau=1}^{m} r_i(t_\tau)$$

Where the $t_\tau$'s are the rounds when we chose action $i$

Now we fix $m$ and get:

$$\forall i \forall m \quad \Pr\left[\left|\hat{\mathbb{V}}_m(i) - \mu_i\right| \leq \sqrt{\frac{8 \log T}{m}}\right] \geq 1 - \frac{2}{T^4}$$

and notice that $\hat{\mu}_t(i) \equiv \hat{\mathbb{V}}_i(m)$ when $m = n_i(t)$.

Define the "good event" $G$:

$$G = \{\forall_i \forall_t \, |\hat{\mu}_i(t) - \mu_i| \leq \lambda_i(t)\}$$

The probability of $G$ is,

$$Pr\,(G) \geq 1 - \frac{2}{T^2}$$

## 15.4 Refine Confidence Bound

Define the upper confidence bound:

$$UCB_t(i) = \hat{\mu}_t(i) + \lambda_t(i)$$

and similarly, the lower confidence bound:

$$LCB_t(i) = \hat{\mu}_t(i) - \lambda_t(i)$$

if $G$ happened then:

$$\forall i \forall t \quad \mu_i \in [LCB_t(i), UCB_t(i)]$$

Therefore:

$$Pr\left[\forall i \forall t \quad \mu_i \in [LCB_t(i), UCB_t(i)]\right] \geq 1 - \frac{2}{T^2}$$

### 15.4.1 Successive Elimination

We maintain a set of actions S.

Initially $S = A$

In each phase:

- We try every $i \in S$ once

- For each $j \in S$ if there exists $i \in S$ such that:

  $$UCB_t(j) < LCB_t(i)$$

  We remove $j$ from $S$, that is we update:

  $$S \leftarrow S - \{j\}$$

We will get the following results:

- As long as action $i$ is still in $S$, we have tried action $i$ exactly the same number of times as all of any other action $j \in S$.
- The best action, under the assumption that the event $G$ holds, is never eliminated from $S$.

Under the assumption of $G$ we get:

$$\mu^* - 2\lambda \leq \hat{\mu}^* - \lambda = LCB_* < UCB_i = \hat{\mu_i} + \lambda \leq \mu_i + 2\lambda$$

Where $\lambda = \lambda_i = \lambda^*$ because we have chosen action $i$ and the best action the same number of times so far.

Therefore, assuming event $G$ holds,

$$\Delta_i = \mu^* - \mu_i \leq 4\lambda = 4\sqrt{\frac{8 \log T}{n_t(i)}}$$

$$\Rightarrow \quad n_T(i) \leq \frac{c}{\Delta_i^2} \log T$$

This implies that

$$\begin{aligned} E\left[\text{Pseudo Regret}\right] &= \sum_{i=1}^{k} \Delta_i n_i(t) \\ &\leq \sum_{i=1}^{k} \frac{c}{\Delta_i} \log T + \underbrace{\frac{2}{T^2} \cdot T}_{\text{The bad event}} \end{aligned}$$

meaning that the expected pseudo regret is bounded by $O\left(\frac{1}{T}\right)$.

### 15.4.2 Upper confidence bound (UCB)

The UCB algorithm simply uses the UCB bound. The algorithm works as follows:

- We try each action once (for a total of $k$ rounds)
- Afterwards we choose:

$$a_t = \arg\max_i UCB_t(i)$$

If we chose action $i$ then, assuming $G$ holds, we have

$$UCB_t(i) \geq UCB_t(a^*) \geq \mu^*$$

where $a^*$ is the optimal action.

Using the definition of UCB and the assumption that $G$ holds, we have

$$UCB_t(i) = \hat{\mu}_t(i) + \lambda_t(i) \leq \mu_i + 2\lambda_t(i)$$

Since we selected action $i$ at time $t$ we have

$$\mu_i + 2\lambda_t(i) \geq \mu^*$$

Rearranging, we have,

$$2\lambda_t(i) \geq \mu^* - \mu_i = \Delta_i$$

Each time we chosen action $i$, we could not have made a very big mistake because:

$$\Delta_i \leq \ 2 \cdot \sqrt{\frac{8 \log T}{n_t(i)}}$$

And therefore if $i$ is very far off from the optimal action we would not choose it too many times. We can bound the number of time action $i$ is used by,

$$n_t(i) \leq \frac{c}{\Delta_i^2} \log T$$

And over all we get:

$$\begin{aligned} E\left[\text{Pseudo Regret}\right] &= \sum_{i=1}^{k} \Delta_i E\left[n_t(i)\right] + \underbrace{\frac{2}{T^2} \cdot T}_{\text{The bad event}} \\ &\leq \sum_{i=1}^{k} \frac{c}{\Delta_i} \cdot \log T + \frac{2}{T} \end{aligned}$$

## 15.5 Best Arm Identification

We would like to identify the best action, or an almost best action. We can define the goal in one of two ways.

**PAC criteria** An action $i$ is $\epsilon$-optimal if $V$. The PAC ctriteria is that, given $\epsilon, \delta > 0$,, with probability at least $1-\delta$, find an $\epsilon$ optimal action.

**Exact identification** Given $\Delta \leq \mu^* - \mu_i$ (for every suboptimal action $i$), find the optimal action $a_*$, with probability at least $1-\delta$.

### 15.5.1 Naive Algorithm (PAC criteria):

We sample each action $i$ for $m = \frac{8}{\epsilon^2}\log\frac{2k}{\delta}$ times, and return $a = \arg\max_i \hat{\mu}_i$ .

For rewards in $[0,1]$, then, by Theorem 15.1, for every action $i$ we have

$$Pr\left[\underbrace{|\hat{\mu}_i - \mu_i| > \frac{\epsilon}{2}}_{\text{bad event}}\right] \leq 2e^{-\left(\frac{\epsilon}{2}\right)^2 m/2} = \frac{\delta}{k}$$

By union bound we get:

$$Pr\left[\exists_i\,|\hat{\mu}_i - \mu_i| > \frac{\epsilon}{2}\right] \leq \delta$$

If the bad event $B = \{\exists_i\,|\hat{\mu}_i - \mu_i| > \frac{\epsilon}{2}\}$ did not happen, then both: (1) $\mu^* - \frac{\epsilon}{2} \leq \hat{\mu}^*$ and (2) $\mu_i + \frac{\epsilon}{2} \leq \hat{\mu}_i$.

This implies,

$$\Rightarrow\ \mu_i + \frac{\epsilon}{2} \geq \hat{\mu}_i \geq \hat{\mu}^* \geq \mu^* - \frac{\epsilon}{2}$$

$$\Rightarrow\ \epsilon \geq \mu^* - \mu_i$$

And therefore $a = \arg\max_i \hat{\mu}_i$ is the optimal action in probability $1-\delta$

We would like to slightly improve the sample size of this algorithm

### 15.5.2 Median Algorithm

The idea: the algorithm runs for $l$ phases, after each phase we eliminate half of the actions. This elimination allows us to sample each action more times in the next phase which makes eliminating the optimal action less likely.

```
**Input**: $\epsilon, \delta > 0$
**Output**: $\bar{a} \in A$
**Init**: $S_1 = A$, $\epsilon_1 = \frac{\epsilon}{4}$, $\delta_1 = \frac{\delta}{2}$, $l = 1$
**Repeat**:
    $\forall_i \in S_l$, sample action i, $\frac{1}{\left(\frac{\epsilon_l}{2}\right)^2} \log\left(\frac{3}{\delta_l}\right)$ times
    $\hat{\mu}_i \leftarrow$ mean (only of samples during the $l^{th}$phase)
    median$_l \leftarrow median\left\{\hat{\mu}_i : i \in S_l\right\}$
    $S_{l+1} \leftarrow \left\{i \in S_l : \hat{\mu}_i \geq \text{median}_l\right\}$
    $\epsilon_{l+1} \leftarrow \frac{3}{4}\epsilon_l$
    $\delta_{l+1} \leftarrow \frac{\delta_l}{2}$
    $l \leftarrow l + 1$
**Until** $|S_l| = 1$
```

**Algorithm 1:** Best Arm Identification

**Complexity:** During phase $l$ we have $|S_l| = \frac{k}{2^{l-1}}$ actions.

$$\epsilon_l = \frac{3}{4}\epsilon_{l-1} = \frac{\epsilon}{4}\left(\frac{3}{4}\right)^{l-1}, \ \delta_l = \frac{\delta}{2^l}$$

$$\Rightarrow \sum \epsilon_l \leq \epsilon, \ \sum \delta_l \leq \delta$$

The total number of samples is therefore:

$$ccc\sum_l |S_l| \cdot \frac{4}{\epsilon_l^2}\log\frac{3}{\delta_l} = \sum_l \frac{k}{2^{l-1}}\frac{64}{\epsilon^2}\left(\frac{16}{9}\right)^{l-1}\log\frac{3 \cdot 2^l}{\delta}$$

$$= \sum_l k\left(\frac{8}{9}\right)^{l-1}\left[c \cdot \frac{\log\frac{1}{\delta}}{\epsilon^2} + \frac{\log 3}{\epsilon^2} + \frac{l}{\epsilon^2}\right]$$

$$= O\left(\frac{k}{\epsilon^2}\log\frac{1}{\delta}\right)$$

**Correctness:**

**Theorem 15.2.** $Pr\left[\underbrace{\max_{j \in S_l}\mu_j}_{action\ l} \leq \underbrace{\max_{j \in S_{l+2}}\mu_j}_{action\ l+1} + \epsilon_l\right] \geq 1 - \delta_l$

*Proof.* We do the proof for $l = 1$, general $l$ is similar. Define $E_1 = \left\{\hat{\mu}^* < \mu^* - \frac{\epsilon_1}{2}\right\}$. We have $Pr\left[E_1\right] \leq \frac{\delta_1}{3}$. If $E_1$ did not happen, we define a bad set:

$$\text{Bad} = \{j \ : \ \mu^* - \mu_j \geq \epsilon_1, \ \hat{\mu}_j \geq \hat{\mu}^*\}$$

Consider an action $j$ such that $\mu^* - \mu_j \geq \epsilon_1$, then:

$$Pr[\hat{\mu}_j \geq \hat{\mu}^* | \underbrace{\hat{\mu}^* \geq \mu^* - \frac{\epsilon_1}{2}}_{\neg E_1}] \leq Pr[\hat{\mu}_j \geq \mu_j + \frac{\epsilon_1}{2} | \neg E_1] \leq \frac{\delta_1}{3}$$

Note that the probability is not negligible. We will show that it cannot happen to too many such actions. We will bound the expectation of the size of *Bad*,

$$E[|\text{Bad}| | \neg E_1] \leq k\frac{\delta_1}{3}$$

with Markov's inequality we get:

$$Pr\left[|\text{Bad}| \geq \frac{k}{2} \middle| \neg E_1\right] \leq \frac{E\,|\text{Bad}|}{k/2} = \frac{2}{3}\delta_1$$

with probability $1 - \delta_1$: $\hat{\mu}^* \geq \mu^* - \frac{\epsilon_1}{2}$ and $|\text{Bad}| \leq \frac{k}{2}$. Therefore: $\exists_j \notin \text{Bad}$ and $j \in S_{l+1}$. $\square$

# Chapter 16

# POMDP

The lecture today is about Partially Observable MDP (POMDP). The main difference is that we will not observe the current state but only a signal that depends on it.

**Example:**
Assume we have four states, in linear order: $1, 2, 3, 4$. We have an action UP that with probability 0.90 moves from $i$ to $i-1$ and with probability 0.1 moves to $i+1$. Similarly, we have an action DOWN that with probability 0.90 moves from $i$ to $i+1$ and with probability 0.1 moves to $i-1$. (In state $i=1$ if we need to move to $i-1$ we stay in $i=1$ and in state $i=4$ if we need to move to $i+1$ we stays in state $i=4$.)

One of the states, state 2 has a reward, while the other states do not have any reward. (See Figure 16.1.) The only observation we see is the reward.

Initially, we know we are in a random state without a reward. This implies that the distribution of where we are is $(1/3, 0, 1/3, 1/3)$. Assume we did an action UP and did not observe a reward. We now can compute the posterior of our probability distribution.

First, we compute the state probability distribution, assuming any current state.

1. Assuming we are in state $i=1$, our state distribution will be $(0.9, 0.1, 0, 0)$.

2. Assuming we are in state $i=2$, our state distribution will be $(0.9, 0, 0.1, 0)$.

3. Assuming we are in state $i=3$, our state distribution will be $(0, 0.9, 0, 0.1)$.

4. Assuming we are in state $i=4$, our state distribution will be $(0, 0, 0.9, 0.1)$.

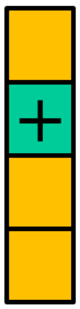


Figure 16.1: The four state POMDP

Since our prior state distribution was $(1/3, 0/1/3, 1/3)$ our posterior state distribution (before observing the reward) is $(0.3, 1/3, 0.3, 1/15)$.

Given that we observed "no reward" we are not instate 2 so the posterior state distribution is $(0.45, 0, 0.45, 0.1)$.

## 16.1 POMDP: model

The model of a POMDP will be similar to that of an MDP with an additional observation function. Namely,

- $S$, a finite set of states.
- $A$, a finite set of actions.
- $p(s'|s, a)$, a probability distribution of next state given that we are in state $s$ and do action $a$.
- $R(s, a)$ a random variable for the reward in state $s$ doing action $a$, and $r(s, a) = E[R(s, a)]$.
- $q_0$ the initial state (can also be a distribution over states)
- $Ob$ is an observability distribution over $O$ a set of observable signals. $O$ can be either finite of infinite. $Ob(o|s', a)$ is the probability that we observe signal $o \in O$ if we reach state $s'$ after doing action $a$. (Note that in an MDP we have the observable signals $O = S$ and $Ob(o|s', a) = s'$).

### 16.1.1 Belief state

The belief state tracks the state distribution, which captures our current state. Each time we perform an action and observe an observation we compute a new belief state, which is our posterior distribution, given our action and observation.

The belief state is a sufficient statistics, implying that it has all the information needed from the history. (Also, two histories which reach the same belief state, we should behave identical for them.) Formally, a belief state $b$ is a distribution over states, i.e., $b \in [0,1]^{|S|}$ and $\sum_s b(s) = 1$.

Given a belief state $b$ and an action $a$ and observation $o$, we need to compute the next belief state $b'(s) = \Pr[s'|o,a,b]$. This will be a deterministic function, and we will have $b' = T(b,a,o)$.

**Belief state computation:**
We compute the next belief state as follows:

$$\begin{aligned}
b'(s') &= \Pr[s'|o,a,b] \\
&= \frac{\Pr[o|s',a,b]\Pr[s'|a,b]}{\Pr[o|a,b]} \\
&= \frac{\Pr[o|s',a]\sum_s \Pr[s'|a,b,s]\Pr[s|a,b]}{\Pr[o|a,b]} \\
&= \frac{Ob(o|s',a)\sum_s p(s'|a,s)b(s)}{\Pr[o|a,b]}
\end{aligned}$$

where the first identity is the definition of the new belief state. The second identity is Bayes rule. The third identity is adding a conditioning over $s$. The last identity is translating the probabilities to the POMDP parameters.

**From POMDP to infinite MDP**
Consider the following MDP. The set of states are $B$ all the possible belief states. (Note that $B$ is infinite!) The set of actions is $A$ (finite). For each belief state $b$ and action $a$ we define the expected reward $r(b,a) = \sum_s b(s)r(s,a)$. The transition probability is to move to $b' = T(b,a,o)$ and is deterministic. (Namely, $\Pr[b' = T(b,a,o)|b,a,o] = 1$ and $\Pr[b' \neq T(b,a,o)|b,a,o] = 0$.)

## 16.2 Value function

The value function will be mapping belief states to expected return. The return can be any of the return we discussed: finite horizon, discounted, of average reward.

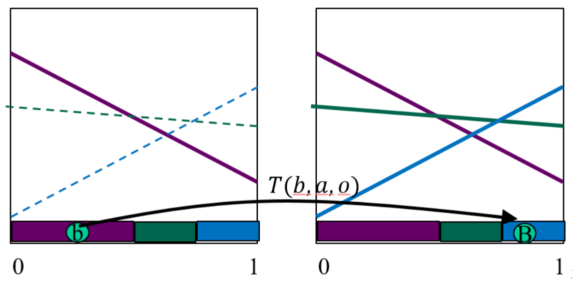


Figure 16.2: Computing $V_2^*(b|a_1, o_1)$

We would like to characterize the optimal value function. For a finite horizon it will be piece-wise linear and convex. (For discounted it will be convex.) Our algorithmic challenge is to handle the infinite state spaces. We will concentrate on the case of finite horizon.

For simplicity, we start with a horizon of length $H = 1$. Namely, we would like to maximize our immediate reward. For each action, our immediate reward is a linear function in the belief state. Namely, $r(b, a) = \sum_s b(s)r(s, a)$. The optimal action for $H = 1$ is the action $a$ that maximizes the expected reward. Namely, $V_1^*(b) = \max_a r(b, a) = \max_a[\sum_s b(s)r(s, a)]$. This implies that the optimal value function is a maximum of linear functions. This gives us a finite representations. It also partitions the space of belief states to regions, which in each region there is an optimal action. The partition to the regions is done by intersecting hyper-planes. The region in which $a_1$ is the best action is define by $\{b : \forall a \neq a_1 \sum_s b(s)r(s, a_1) \geq \sum_s b(s)r(s, a)\}$.

We will now consider a horizon of $H = 2$. We will partition the task to three steps. In the first step, we will assume that we are in belief state $b$ perform action $a_1$ and observe $o_1$. In the second step, we assume we are in belief state $b$ and perform action $a_1$. In the third step, we only assume we are in belief state $b$.

Assume we we are in belief state $b$ perform action $a_1$ and observe $o_1$. We would like to compute $V_H(b|a_1, o_1)$ for $H = 2$. For the first step, given that we do action $a_1$, we can compute the expected immediate reward $\sum_s b(s)r(s, a_1)$. Now, we move from belief state $b$ to belief state $b' = T(b, a_1, o_1)$. When we reach $b'$ we have only one step remaining, so we perform the best action and get $V_1^*(b')$. This implies that $V_2^*(b|a_1, o_1) = r(b, a_1) + V_1^*(b')$. (See Figure 16.2.) This implies that $V_2^*(b|a_1, o_1)$ is piece-wise linear.

We can now do the same for each possible observations (assume the set of observations is finite, for simplicity). This implies that we have the values of $V_2^*(b|a_1, o)$ for any $o \in O$. This allows us to compute $V_2^*(b|a_1)$, by simply averaging over the observations. $V_2^*(b|a_1) = \sum_o V_2^*(b|a_1, o) \Pr[o|b, a_1]$. We can compute $\Pr[o|b, a] = \sum_{s,s'} b(s)p(s'|s, a)Ob(o|s', a)$.

Given that we can compute $V_2^*(b|a)$ for any action $a \in A$, we can deduce $V_2^*(b) = \max_a V_2^*(b|a)$.

We can now compute for a longer horizon, say $H = 3$, using dynamic programming. We want to compute $V_3^*(b)$. We first compute the function $V_2^*(\cdot)$. For each action $a_i$ and observation $o_j$ we compute $V_3^*(b|a_i, o_j)$. This is done by first computing $r(b, a_i)$ and $b' = T(b, a_i, o_j)$. Then we compute $V_3^*(b|a_i, o_j) = r(b, a_i) + V_2^*(b')$.

**Optimal value function**

For the finite horizon the optimal value function $V_H^*(b)$ is define as follows,

$$V_H^*(b) = \max_a [r(b, a) + \sum_o \Pr[o|b, a] V_{H-1}^*(T(b, a, o))]$$

and the optimal policy is

$$\pi_H^*(b) = \arg\max_a [r(b, a) + \sum_o \Pr[o|b, a] V_{H-1}^*(T(b, a, o))]$$

**Theorem 16.1.** *For the finite horizon, the function $V_H^*$ is piece-wise linear and convex. In addition, there exists a collection of $\Theta = \{\theta_i \in \mathbb{R}^{|S|}\}$ such that $V_H^*(b) = \max_{\theta_i \in \Theta} \sum_s b(s)\theta_i(s)$.*

For the discounted setting we have

$$V^*(b) = [r(b, a) + \gamma \sum_o \Pr[o|b, a] V_{H-1}^*(T(b, a, o))]$$

**Theorem 16.2.** *For the discounted return, the function $V^*$ is convex.*

**Policy Tree**

How does the optimal finite horizon policy can be implemented. We can create a tree. In each node of the tree, we can label it by the action we perform. We label the outgoing edges by the observation we receive. We start at the root. The action at the root is the action we perform at the initial state. Given the observation we receive, we continue to the next node in the tree. The label of that node is the action we perform and so on. The depth of the tree is $H$ for a finite horizon of length $H$.

### 16.2.1 Value Iteration:

We can run a value iteration in the POMDP (actually, on the belief states. We can write the update as follows:

$$\begin{aligned}
V_t^{a,o}(b) =& \frac{r(b,a)}{|O|} + \gamma \Pr[o|b,a] V_t(T(b,a,o)) \\
V_t^a(b) =& \sum_o V_t^{a,o}(b) \\
V_{t+1}^*(b) =& \max_a V_t^a(b) \\
V_{t+1}^*(b) =& \max_a [r(b,a) + \gamma \sum_o V_t^*(T(b,a,o)
\end{aligned}$$

Assume that $V_t$ is piece-wise linear. Namely, there exists a set $\Theta_t$ such that

$$V_t(b) = \max_{\theta \in \Theta_t} \theta^\top b$$

For $V_t^{a,o}$ we have that $b' = T(b,a,o)$ is linear in $b$, i.e., there is a matrix $M$ such that $b' = Mb$. (Note that $M$ depends on $a$ and $o$.) The value of $V_t^{a,o}$ is

$$V_t^{a,o}(b') = \max_{\theta \in \Theta_t} \theta^\top b' = \max_{\theta \in \Theta_t} \theta^\top M b$$

This implies that we can define $\Theta_t^{a,o} = \{\theta^\top M : \theta \in \Theta_t\}$.

For $V_t^a$ we have that

$$V_t^a(b) = \sum_o V_t^{a,o}(b) = \sum_o \max_{\theta \in \Theta_t^{a,o}} \theta^\top b = \max_{\theta \in \Theta_t^a} \sum_o \theta^\top b$$

where

$$\Theta_t^a = \{\sum_o \theta^{a,o} | \theta^{a,o} \in \Theta_t^{a,o}\}$$

For $V_{t+1}^*$ we have

$$V_{t+1}^*(b) = \max_a V_t^a(b) = \max_{\theta \in \Theta_{t+1}} \theta^\top b$$

where $\theta_{t+1} = \cup_a \Theta_t^a$.

The complexity of value iteration depends on $\Theta_t$. For each action $a$ and observation $o$ we have $|\Theta_t^{a,o}| \le |\Theta_t|$. For each action $a$ we have $|\Theta_t^a| \le |\Theta_t^{a,o}|^{|O|} \le |\Theta_t|^{|O|}$. For $\Theta_{t+1}$ we have $|\Theta_{t+1}| \le \sum_a |\Theta_t^a| \le |A| \cdot |\Theta_t|^{|O|}$.

The complexity is exponential in $|O|$ in each iteration! Pruning can help, but the exponential time seems unavoidable.

### 16.2.2 Hardness

Many computational problems related to POMDPs are hard. For the finite horizon, even with a unique start state, it is P-SPACE complete to compute the return of the optimal policy.

For the finite horizon, in the case that there are no observations, it is NP complete to compute the return of the optimal policy. (We will show this simple hardness result.)

Assume we have no observations. This implies that we need to compute a sequence of $H$ actions. The optimal policy will be the sequence of $H$ actions that will maximize the return. We will show a reduction to SAT.

Given a SAT formula over $n$ variables we create the following POMDP. For each clause we create the following POMDP gadget. We have always two actions 0 and 1. We create two lines of length $n$, one line ends in a state with reward 1 and the other in a state with reward 0. All other rewards will be zero. We "think" of the state in the line as a number in $[1, n]$ and we associate the state $i$ it with the variable $x_i$.

We start at the line ending with 0. All the transition of states in the line ending in 1 simply continue to the next state, i.e., from $i$ to $i+1$ (for both actions). For the line ending in zero, for each state $i$ where $x_i$ nor $\bar{x}_i$ do not appear in the clause, we continue to the next state on this line, with both actions.

For state $i$ in the line that ends in 0, if $x_i$ appears in the clause, then in state $s_i$ with action 1 we move to the $i+1$ state in the line ending in 1, and with action 0 we move to state $i+1$ in the line ending in zero. Similarly, if $\bar{x}_i$ appears in the clause, then in state $s_i$ with action 0 we move to the $i+1$ state in the line ending in 1, and with action 1 we move to state $i+1$ in the line ending in zero.

It is not hard to see that the sequence of $n$ actions will lead to the reward 1 if and only if it represents an satisfying assignment to the clause.

We now take the gadgets for all the clauses, and select our initial state to uniformly at random to be the start state of one of the gadget (which represent a clause).

If the SAT formula is satisfiable, there is an assignment that satisfies all the clauses. Therefore guarantees a return of 1.

Otherwise for each sequence of action, which represent an assignment, there is some clause which is not satisfied. This implies that the return is at most $1 - 1/m$, where $m$ is the number of clause.

**Remark:** Actually, we can get from the same reduction also a hardness to approximate result. We simply consider the MAX-SAT problem. For 3-SAT it is NP-hard to decide distinguish between satisfying $7/8 + \epsilon$ of the terms or all the terms. This implies that it is NP-hard to distinguish between a return of $7.8 + \epsilon$ or 1.

### 16.2.3 Policy Plan: Automata

A simple class to describe the policy is a finite automata with output, which is called a Moore automata. The outputs will be the selected actions and the inputs are the observations. While it is true that this class of policies is not universal, namely, not any policy can be represented in this way, many policies can be.

One nice observation regarding this class of polices is that we can efficiently compute the value function of a given automata. Assume that $\pi$ is described by an automata. The value of $\pi$ from state $s$ when its automata is in state $\sigma$ is

$$V^{\pi}(s;\sigma) = r(s,a) + \gamma \sum_{o,s'} p(s'|s,a)Ob(o|s'a)V^{\pi}(s';\sigma')$$

where $a = \pi(\sigma, o)$ the action selected by $\pi$ when in state $\sigma$ of the automata and observes $o$, and $\sigma'$ is the next state in the automata, given we are in state $\sigma$ and observe $o$.

This implies that we have a set of linear equations and we can solve them efficiently!

## 16.3 Reusable trajectories

Our goal is given a set of deterministic policies $\Pi$, to estimate for each $\pi \in \Pi$ the value $V^{\pi}(s_0)$, its expected return. We want a set of trajectories that would be relatively small compared to $\Pi$, hopefully logarithmic in $|\Pi|$.

We will generate a set of trajectories $\Gamma$ in the following simple way. We use a completely random policy to select the actions. Namely, $\pi_R(a|b) = 1/|A|$. We generate $m$ such trajectories.

We estimate the return of a policy $\pi \in \Pi$ in the following way. Let $\Gamma_\pi$ all the trajectories that agree with $\pi$. (Recall, $\pi$ is deterministic.) We set

$$\hat{V}^{\pi}(s_0) = \frac{1}{|\Gamma_\pi|} \sum_{h \in \gamma_\pi} return(h)$$

The main issue will be: how large $m$ has to be, in order to guarantee with probability $1-\delta$ that the error for any $\pi \in \Pi$ in the estimation will be at most $\epsilon$. We will look at $m$ as a function of the error $\epsilon$, the confidence $\delta$, the number of policies $|\Pi|$, and the maximum return $V_{max}$.

The proof follows in the following steps. Let $acc_\pi(h) = 1$ if $\pi$ agrees on all the actions in $h$. The following lemma states that for any policy, the probability that it agrees with the trajectory of length $H$ is exactly $1/|A|^H$.

**Lemma 16.3.** *For a finite horizon $H$, for any policy $\pi$ we have* $\Pr_h[acc_\pi(h) = 1] = \frac{1}{|A|^H}$.

Let $D_\pi$ be the distribution over histories generated by policy $\pi$. The following lemma states that the distribution of the random policy, condition of policy $\pi$ agreeing with the actions, is identical to that of generated by policy $\pi$.

**Lemma 16.4.**
$$D_\pi(h) = D_{\pi_R}(h|acc_\pi(h) = 1)$$

The following lemma states that with high probability, for every $\pi \in \Pi$ we have that the size of $|\Gamma_\pi|$ is exponential in the horizon.

**Lemma 16.5.** *For*
$$m > 8|A|^H \frac{V_{max}^2}{\epsilon^2} \log(2|\Pi|/\delta)$$
*with probability $1 - \delta/2$ for every $\pi \in \Pi$ we have*
$$|\Gamma_\pi| \geq \frac{m}{2^{H+1}} > 4\frac{V_{max}^2}{\epsilon^2} \log(2|\Pi|/\delta)$$

The following lemma states that if every $\Gamma_\pi$ is of sufficient size, then for each policy $\pi$ we have a good approximation.

**Lemma 16.6.** *If for every $\pi \in \Pi$ we have*
$$|\Gamma_\pi| \geq \frac{m}{2|A|^H} > 4\frac{V_{max}^2}{\epsilon^2} \log(2|\Pi|/\delta)$$
*then with probability $1 - \delta/2$, for every $\pi \in \Pi$ we have*
$$|\hat{V}^\pi(s_0) - V^\pi(s_0)| \leq \epsilon$$

The theorem concludes and establishes the sample size for a PAC guarantee.

**Theorem 16.7.** *For*
$$m > 8|A|^H \frac{V_{max}^2}{\epsilon^2} \log(2|\Pi|/\delta)$$
*with probability $1 - \delta$, for every $\pi \in \Pi$ we have*
$$|\hat{V}^\pi(s_0) - V^\pi(s_0)| \leq \epsilon$$

# Chapter 17

# LQR

# Chapter 18

# Inverse RL and Apprenticeship learning

## 18.1 Inverse Reinforcement Learning

The classical reinforcement problem has as inputs the dynamics and the rewards and the output is the optimal policy. In the inverse reinforcement learning problem we are given access to the optimal policy and the dynamics, and would like to learn a reward function which induces the given optimal policy.

More specifically, we are given as inputs the states $S$, actions $A$, dynamics $p(\cdot|s, a)$, and access to the optimal policy $\pi^*$, either explicitly or implicitly, by observing trajectories of $\pi^*$.

We have a few related problems:

1. Behavioral cloning: given trajectories learn the optimal policy.

2. Inverse Reinforcement Learning (IRL): which recovers a reward function $R$ for which $\pi^*$ is optimal.

3. Apprentice learning: Learn a near-optimal policy given access to trajectories of an optimal policy.

### 18.1.1 Behavioral Cloning

Formulated as a classical supervised classification problem. Given trajectories generated from $\pi^*$, we extract pairs $(s_t, a_t)$, where $\pi^*(s_t) = a_t$. We do not need to know the dynamics or the rewards. We define a classification problem, where the training examples are $\mathcal{S} = \{(s_t, a_t)\}$.

We can now learn a classifier for $\mathcal{S}$, and use any classifier, DNN, SVM, decision tree, and more. We learn the optimal policy as a classification problem, and if we have low error we can hope for performance near that of the optimal policy.

There are a few problems with the approach. After we make errors we might reach completely new states, which we never encountered before (since the optimal policy never reaches them). Also, the different errors might have very different effects, but we treat them all equally. Other issues include generalization, which would depend on the test and train distributions (of states) being identical, but small differences in the policies might lead to significant differences in the distributions.

### 18.1.2 IRL: small state space

Assume we can use a look-up table to represent the policies. We are given the set of states $S$, actions $A$, $p(\cdot|s,a)$ and the optimal policy $\pi^*$. (All are given explicitly, as look-up tables.) The output of the algorithm is a reward function $r(s,a)$ for which $\pi^*$ is optimal.

We first characterize all the rewards $R$ for which $\pi^*$ is optimal. Bellman optimality requires that for every $s \in S$ and $a \in A$ we have,

$$Q^*(s, \pi^*(s)) \geq Q^*(s,a)$$

First we will try to get a more explicit characterization.

For simplicity of the notation let $a_1 = \pi^*(s)$ for every $s \in S$. (This is simply renaming the actions.) Also, we associate the rewards with states (rater then state-action pairs). Define matrices $P^a$, for $a \in A$, as follows: $P^a[i,j] = p(s_j|s_i, a)$, which is the transition probabilities using action $a$. Since $a_1$ is the optimal action, we define $P^* = P^{a_1}$.

For the optimal value function $V^* \in \mathbb{R}^{|S|}$ has

$$V^* = R + \gamma P^* V^*$$

This implies that $V^* = (I - \gamma P^*)^{-1} R$. (The inverse exists, since the eigenvalues of $P^*$ are in the unit circle, hence, all the eigenvalues of $I - \gamma P^*$ are non-zero.)

The optimal state-action function is,

$$Q^* = R + \gamma P^a V^* = R + \gamma P^a (I - \gamma P^*)^{-1} R$$

From the Bellman optimality we have $Q^*(s, \pi^*(s)) - Q^*(s,a) \geq 0$ which implies,

$$(P^* - P^a)(I - \gamma P^*)^{-1} R \geq 0$$

This establishes the following claim:

**Claim 18.1.** *Reward function $R$ is optimal iff* $(P^* - P^a)(I - \gamma P^*)^{-1} R \geq 0$.

Although we have an exact characterization, we have a few challenges.

1. Trivial solutions: We can simply set $R = 0$ and any policy is optimal!
2. What happens if we observe only trajectories of $\pi^*$ and not the explicit policy and dynamics.
3. Large state space.

To overcome the ambiguity of the reward function we can set an objective that will force a unique outcome. One solution is a *max margin*, specifically,

$$\max_{s \in S} \left\{ Q^*(s, a^*) - \max_{a \neq a^*} Q^*(s, a) \right\}$$

The resulting linear program is,

$$\begin{aligned} & \max \lambda \\ \forall a^* \neq a \quad & (P^* - P^a)(I - \gamma P^*)^{-1} R \geq \lambda \\ & -R_{max} \leq R \leq R_{max} \end{aligned}$$

### 18.1.3 AL: Linear function approximation

In this section we consider the case where we are given access to trajectories of the optimal policy, in a large state space. Our goal is to find a near optimal policy. We consider a simple case that the rewards are linear in the state representation. For each state $s \in S$ we have $\phi(s) \in \mathbb{R}^n$. There is a reward weights $w \in \mathbb{R}^n$. The reward in state $s$ is $w^\top \phi(s)$.

Consider the value function of the policy $\pi$.

$$\begin{aligned} V^*(s_0) =& E[\sum_{t=0}^{\infty} \gamma^t r_t | \pi] \\ =& E[\sum_{t=0}^{\infty} \gamma^t w^\top \phi(s_t) | \pi] \\ =& w^\top \sum_{t=0}^{\infty} \gamma^t E[\phi(s_t) | \pi] \\ =& w^\top \mu(\pi) \end{aligned}$$

where $\mu(\pi) = \sum_{t=0}^{\infty} \gamma^t E[\phi(s_t)|\pi]$. Note that $\mu(\pi)$ depends on the dynamics (steady state distribution) that policy $\pi$ induces.

For IRL we like to recover a weight vector $w$ such that for any policy $\pi$ we have,

$$w^\top \mu(\pi^*) \geq w^\top \mu(\pi)$$

First, we can approximate $\mu(\pi)$ from sample of $m$ trajectories,

$$\mu(\pi) \approx \hat{\mu}(\pi) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{H} \gamma^t \phi(s_t)$$

Second, observe that if we have $\|\mu(\pi) - \hat{\mu}(\pi)\|_2 \leq \epsilon$ then for any $w$ we have

$$|w^\top \mu(\pi) - w^\top \hat{\mu}(\pi)| \leq \epsilon \|w\|_2$$

For the AL we like to find a near optimal policy $\pi'$, but we do not have access to the rewards. However, if we find a policy $\pi'$ such that $\mu(\pi') \approx \mu(\pi^*)$, then for any weights $w$ the policy $\pi'$ will give a similar expected return as $\pi^*$.

Consider first the IRL problem. As before, we have a problem that $w = 0$ is a solution. We will overcome this by defining a max margin linear program.

$$\begin{array}{ll} & \max \lambda \\ \forall \pi & w^\top \mu(\pi^*) \geq w^\top \mu(\pi) + \lambda \\ & \|w\|_2^2 \leq 1 \end{array}$$

A clear weakness is that we have an exponential size linear program. We will overcome this by an *incremental constraint generation* which will solve the AL. We maintain a set of policies $\Pi$, where initially $\Pi = \emptyset$. Initially we select some $\pi^0$ and let $\mu^0 = \mu(\pi^0)$. At iteration $i$ we solve

$$\max_{w} \min_{\pi_j \in \Pi} w^\top (\mu^* - \mu^j)$$

If the value is at most $\epsilon$ we terminate. Otherwise let $w^i$ be the maximizer. We solve for the optimal policy $\pi^i$ given weights $w^i$ and add $\pi^i$ to $\Pi$. At termination, we compute a $\mu = \sum_{i=1}^{k} a_i \mu^i$ which minimizes $\|\mu^* - \mu\|_2$, where $a_i \geq 0$ and $\sum_i a_i = 1$.

The correctness of the constraint generation follows from the following lemma.

**Lemma 18.2.** *At termination* $\|\mu^* - \mu\|_2 \leq \epsilon$

We can implement $\mu$ by selecting policy $\pi^i \in \Pi$ with probability $a_i$ and running it to completion.

The main issue is the rate of convergence. How long will it take to terminate. The following lemma guarantees the fast convergence.

**Lemma 18.3.** *The number of iterations is bounded by*

$$O(\frac{n}{(1-\gamma)^2\epsilon^2}\log\frac{n}{(1-\gamma)\epsilon})$$

## 18.2 Bibliography Remarks

The inverse reinforcement learning follows [1, 10]

# Bibliography

[1] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, 2004.

[2] Ronen I. Brafman and Moshe Tennenholtz. R-MAX - A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

[3] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2818–2826, 2015.

[4] Eyal Even-Dar and Yishay Mansour. Learning rates for q-learning. *Journal of Machine Learning Research*, 5:1–25, 2003.

[5] Michael J. Kearns and Satinder P. Singh. Finite-sample convergence rates for q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems 11, [NIPS Conference, Denver, Colorado, USA, November 30 - December 5, 1998]*, pages 996–1002, 1998.

[6] Michael J. Kearns and Satinder P. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

[7] Nate Kohl and Peter Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation, ICRA 2004, April 26 - May 1, 2004, New Orleans, LA, USA*, pages 2619–2624, 2004.

[8] Lihong Li. Sample complexity bounds of exploration. In Marco Wiering and Martijn van Otterlo, editors, *Reinforcement Learning: State-of-the-Art*, pages 175–204. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[9] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[10] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 663–670, 2000.

[11] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[12] Satinder P. Singh and Richard S. Sutton. Reinforcement learning with replacing eligibility traces. *Machine Learning*, 22(1-3):123–158, 1996.

[13] Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite mdps: PAC analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

[14] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction.* Adaptive computation and machine learning. MIT Press, 1998.

[15] Gerald Tesauro. Temporal difference learning and td-gammon. *Commun. ACM*, 38(3):58–68, 1995.

[16] Gerald Tesauro. Programming backgammon using self-teaching neural nets. *Artif. Intell.*, 134(1-2):181–199, 2002.

[17] J. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Trans. on Automatic Control*, 42(5):674–690, 1997.

[18] Harm van Seijen, Hado van Hasselt, Shimon Whiteson, and Marco A. Wiering. A theoretical and empirical analysis of expected sarsa. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, ADPRL 2009, Nashville, TN, USA, March 31 - April 1, 2009*, pages 177–184, 2009.