Chapter 0

Notation

The reader might find it helpful to refer back to this notation section.

- For a vector \( \mathbf{v} \), we let \( (\mathbf{v})^2 \), \( \sqrt{\mathbf{v}} \), and \(|\mathbf{v}|\) be the component-wise square, square root, and absolute value operations.
- Inequalities between vectors are elementwise, e.g. for vectors \( \mathbf{v}, \mathbf{v}' \), we way \( \mathbf{v} \leq \mathbf{v}' \), if the inequality holds elementwise.
- For a vector \( \mathbf{v} \), we refer to the \( j \)-th component of this vector by either \( \mathbf{v}(j) \) or \( [\mathbf{v}]_j \).
- Denote the variance of any real valued \( f \) under a distribution \( \mathcal{D} \) as:
  \[
  \text{Var}_\mathcal{D}(f) := E_{x \sim \mathcal{D}}[f(x)^2] - (E_{x \sim \mathcal{D}}[f(x)])^2
  \]
- It is helpful to overload notation and let \( P \) also refer to a matrix of size \( (S \cdot A) \times S \) where the entry \( P_{(s,a),s'} \) is equal to \( P(s'|s,a) \). We also will define \( P^\pi \) to be the transition matrix on state-action pairs induced by a deterministic policy \( \pi \). In particular, \( P^\pi_{(s,a),(s',a')} = P(s'|s,a) \) if \( a' = \pi(s') \) and \( P^\pi_{(s,a),(s',a')} = 0 \) if \( a' \neq \pi(s') \). With this notation,
  \[
  Q^\pi = (1 - \gamma)r + \gamma PV^\pi
  \]
  \[
  Q^\pi = (1 - \gamma)r + \gamma P^\pi Q^\pi
  \]
  \[
  Q^\pi = (1 - \gamma)(I - \gamma P^\pi)^{-1}r
  \]
- For a vector \( Q \in \mathbb{R}^{S \times A} \), denote the greedy policy and value as:
  \[
  \pi_Q(s) := \arg\max_{a \in A} Q(s,a)
  \]
  \[
  V_Q(s) := \max_{a \in A} Q(s,a). .
  \]
- For a vector \( Q \in \mathbb{R}^{S \times A} \), the Bellman optimality operator \( T : \mathbb{R}^{S \times A} \to \mathbb{R}^{S \times A} \) is defined as:
  \[
  TQ := (1 - \gamma)r + \gamma PV_Q. \tag{0.1}
  \]
Chapter 1

MDP Preliminaries
1.1 Markov Decision Processes

In reinforcement learning, the interactions between the agent and the environment are often described by a Markov Decision Process (MDP) [Puterman, 1994], specified by:

- **State space** $S$. In this course we only consider finite state spaces.
- **Action space** $A$. In this course we only consider finite action spaces.
- **Transition function** $P : S \times A \rightarrow \Delta(S)$, where $\Delta(S)$ is the space of probability distributions over $S$ (i.e., the probability simplex). $P(s'|s,a)$ is the probability of transitioning into state $s'$ upon taking action $a$ in state $s$.
- **Reward function** $r : S \times A \rightarrow [0,1]$. $r(s,a)$ is the immediate reward associated with taking action $a$ in state $s$.
- **Discount factor** $\gamma \in [0,1)$, which defines a horizon for the problem.

1.1.1 Interaction protocol

In a given MDP $M = (S,A,P,r,\gamma)$, the agent interacts with the environment according to the following protocol: the agent starts at some state $s_0$; at each time step $t = 0,1,2,\ldots$, the agent takes an action $a_t \in A$, obtains the immediate reward $r_t = r(s_t,a_t)$, and observes the next state $s_{t+1}$ sampled according to $s_{t+1} \sim P(\cdot|s_t,a_t)$. The interaction record at time $t$ $\tau_t = (s_0,a_0,r_1,s_1,\ldots,s_t)$ is called a *trajectory*, which includes the observed state at time $t$.

In some situations, it is necessary to specify how the initial state $s_0$ is generated. We consider $s_0$ sampled from an initial distribution $\mu \in \Delta(S)$. When $\mu$ is of importance to the discussion, we include it as part of the MDP definition, and write $M = (S,A,P,r,\gamma,\mu)$.

1.1.2 The objective, policies, and values

In the most general setting, a policy specifies a decision-making strategy in which the agent chooses actions adaptively based on the history of observations; precisely, a policy is a mapping from a trajectory to an action, i.e. $\pi : \mathcal{H} \rightarrow A$ where $\mathcal{H}$ is the set of all possibly trajectories. A deterministic, stationary policy $\pi : S \rightarrow A$ specifies a decision-making strategy in which the agent chooses actions adaptively based on the current state, i.e., $a_t = \pi(s_t)$. The agent may also choose actions according to a stochastic policy $\pi : S \rightarrow \Delta(A)$, and, overloading notation, we write $a_t \sim \pi(\cdot|s_t)$. A deterministic policy is its special case when $\pi(s)$ is a point mass for all $s \in S$.

For a fixed policy and a starting state $s_0 = s$, we define the value function $V^\pi_M : S \rightarrow \mathbb{R}$ as the average, discounted
sum of future rewards

\[ V^\pi_M(s) = (1 - \gamma) \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid \pi, s_0 = s \right]. \]

where expectation is with respect to the randomness of the trajectory, that is, the randomness in state transitions and the stochasticity of \( \pi \). Here, the factor of \( 1 - \gamma \) serves as a normalizing factor: since \( r(s, a) \) is bounded between 0 and 1, we have \( 0 \leq V^\pi_M(s) \leq 1 \).

Similarly, the action-value (or Q-value) function \( Q^\pi_M: \mathcal{S} \times \mathcal{A} \to \mathbb{R} \) is defined as

\[ Q^\pi_M(s, a) = (1 - \gamma) \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid \pi, s_0 = s, a_0 = a \right]. \]

Given a state \( s \), the goal of the agent is to find a policy \( \pi \) that maximizes the value, i.e. the optimization problem the agent seeks to solve is:

\[ \max_{\pi} V^\pi_M(s) \]

The dependence of on \( M \) may be dropped when it is clear from context.

**Example 1 (Navigation).** Navigation is perhaps the simplest to see example of RL. The state of the agent is their current location. The four actions might be moving 1 step along each of east, west, north or south. The transitions in the simplest setting are deterministic. Taking the north action moves the agent one step north of their location, assuming that the size of a step is standardized. The agent might have a goal state \( g \) they are trying to reach, and the reward is 0 until the agent reaches the goal, and 1 upon reaching the goal state. Since the discount factor \( \gamma < 1 \), there is incentive to reach the goal state earlier in the trajectory. As a result, the optimal behavior in this setting corresponds to finding the shortest path from the initial to the goal state, and the value function of a state, given a policy is \( (1 - \gamma)^d \), where \( d \) is the number of steps required by the policy to reach the goal state.

**Example 2 (Conversational agent).** This is another fairly natural RL problem. The state of an agent can be the current transcript of the conversation so far, along with any additional information about the world, such as the context for the conversation, characteristics of the other agents or humans in the conversation etc. Actions depend on the domain. In the rawest form, we can think of it as the next statement to make in the conversation. Sometimes, conversational agents are designed for task completion, such as travel assistant or tech support or a virtual office receptionist. In these cases, there might be a predefined set of slots which the agent needs to fill before they can find a good solution. For instance, in the travel agent case, these might correspond to the dates, source, destination and mode of travel. The actions might correspond to natural language queries to fill these slots.

In task completion settings, reward is naturally defined as a binary outcome on whether the task was completed or not, such as whether the travel was successfully booked or not. Depending on the domain, we could further refine it based on the quality or the price of the travel package found. In more generic conversational settings, the ultimate reward is whether the conversation was satisfactory to the other agents or humans, or not.

**Example 3 (Board games).** This is perhaps the most popular category of RL applications, where RL has been successfully applied to solve Backgammon, Go and various forms of Poker. For board games, the usual setting consists of the state being the current game board, actions being the potential next moves and reward being the eventual win/loss outcome or a more detailed score when it is defined in the game.

### 1.1.3 Bellman consistency equations for stationary policies

By definition, \( V^\pi \) and \( Q^\pi \) satisfy the following Bellman consistency equations: for all \( s \in \mathcal{S}, a \in \mathcal{A} \),

\[ V^\pi(s) = Q^\pi(s, \pi(s)). \]

\[ Q^\pi(s, a) = (1 - \gamma) r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)} [V^\pi(s')] , \]

(1.2)
where we are treating $\pi$ as a deterministic policy.

It is helpful to view $V_\pi$ as vector of length $S$ and $Q_\pi$ and $r$ as vectors of length $S \cdot A$. We overload notation and let $P$ also refer to a matrix of size $(S \cdot A) \times S$ where the entry $P(s', s, a')$ is equal to $P(s'|s, a)$. We also will define $P_\pi$ to be the transition matrix on state-action pairs induced by a deterministic policy $\pi$. In particular,

$$P_\pi(s, a), (s', a') := \begin{cases} P(s'|s, a) & \text{if } a' = \pi(s') \\ 0 & \text{if } a' \neq \pi(s') \end{cases}$$

For a randomized stationary policy, we have $P_\pi(s, a), (s', a') = P(s'|s, a)\pi(a'|s')$. With this notation, it is straightforward to verify:

$$Q_\pi = (1 - \gamma)r + \gamma PV_\pi$$

(1.3)

$$Q_\pi = (1 - \gamma)r + \gamma P_\pi Q_\pi.$$ (1.4)

The above implies that:

$$Q_\pi = (1 - \gamma)(I - \gamma P_\pi)^{-1}r$$

(1.5)

where $I$ is the identity matrix. To see that the $(I - \gamma P_\pi)$ is invertible, observe that for any non-zero vector $x \in \mathbb{R}^{S \cdot A}$,

$$\| (I - \gamma P_\pi) x \|_{\infty} = \| x - \gamma P_\pi x \|_{\infty}$$

(triangular inequality for norms)

$$\geq \| x \|_{\infty} - \gamma \| P_\pi x \|_{\infty}$$

(each element of $P_\pi x$ is a convex average of $x$)

$$= (1 - \gamma) \| x \|_{\infty} > 0$$

(\gamma < 1, x \neq 0)

which implies $I - \gamma P_\pi$ is full rank.

### 1.1.4 Bellman optimality equations

Due to the Markov structure, there exists a single stationary and deterministic policy that simultaneously maximizes $V^*(s)$ for all $s \in \mathcal{S}$ and maximizes $Q^*(s, a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$ [Puterman, 1994]; we denote this optimal policy as $\pi^*_M$ (or $\pi^*$). We use $V^*$ and $Q^*$ as a shorthand for $V^{\pi^*}$ and $Q^{\pi^*}$, respectively.

$V^*$ and $Q^*$ satisfy the following set of Bellman optimality equations [Bellman, 1956]: for all $s \in \mathcal{S}, a \in \mathcal{A}$,

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a).$$

$$Q^*(s, a) = (1 - \gamma)r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s, a)}[V^*(s')].$$ (1.6)

Let us use shorthand $\pi_Q$ to denote the greedy policy with respect to a vector $Q \in \mathbb{R}^{S \cdot A}$, i.e

$$\pi_Q(s) := \arg\max_{a \in \mathcal{A}} Q(s, a).$$

With this notation, the optimal policy $\pi^*$ is obtained by choosing actions greedily (with arbitrary tie-breaking mechanisms) with respect to $Q$, i.e.

$$\pi^* = \pi_{Q^*}.$$ 

Let us also use the notation to greedily turn a vector $Q \in \mathbb{R}^{S \cdot A}$ into a vector of length $|S|$.

$$V_Q(s) := \max_{a \in \mathcal{A}} Q(s, a).$$
The Bellman optimality operator $T_M \colon \mathbb{R}^{|S \times A|} \to \mathbb{R}^{|S \times A|}$ is defined as follows: when applied to some vector $Q \in \mathbb{R}^{|S \times A|}$,
\[TQ := (1 - \gamma)r + \gamma PV_Q.\] (1.7)

This allows us to rewrite Equation 1.6 in the concise form: $Q^* = TQ^*$, i.e. $Q^*$ is a fixed point of the operator $T$. The classic result of [Bellman, 1956] shows that if $Q$ satisfies $Q = TQ$, then $Q = Q^*$. We state the result below formally.

**Theorem 1.** Let $Q^*(s, a) = \max_{\pi \in \Pi} Q^\pi(s, a)$ where $\Pi$ is the space of all (non-stationary and randomized) policies. We have that

- There exists a stationary and deterministic policy $\pi$ such that $Q^\pi = Q^*$
- A vector $Q \in \mathbb{R}^{|S \times A|}$ is equal to $Q^*$ if and only if it satisfies $Q = TQ$.

**Proof:** First observe that:

\[
Q^*(s, a) = (1 - \gamma)E \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right]
\]

\[
= (1 - \gamma)E \left[ r(s_0, a_0) + \ldots + r(s_{\tau-1}, a_{\tau-1}) + \gamma^\tau E \left[ \sum_{t=\tau}^{\infty} \gamma^t r(s_{t+\tau}, a_{t+\tau}) \mid s_0 = s, a_0 = a \right] \right]
\]

\[
= (1 - \gamma)E \left[ r(s_0, a_0) + \ldots + r(s_{\tau-1}, a_{\tau-1}) + \gamma^\tau \max_{\pi} \left( E \left[ \sum_{t=0}^{\infty} \gamma^t r(s_{t+\tau}, a_{t+\tau}) \mid \pi, s_0 = s, a_0 = a \right] \right) \right].
\]

where the inner max is over all policies which may also use the history of information before time $\tau$. Note that $s_\tau$ and $a_\tau$ the future evolution at time $\tau$ does not depend on the $(s_0, a_0, \ldots, s_{\tau-1}, a_{\tau-1})$, which implies that the max value can be achieved with a policy that, at time $\tau$, chooses an action that only depends on $s_\tau$. This proves the stationarity claim. Furthermore, by linearity of expectation, the choice of $a_\tau$ can made deterministically.

For the second claim, we first show that $Q^*$ satisfies $Q^* = TQ^*$. We need only consider deterministic policies. We have:

\[
Q^*(s, a) = \max_\pi Q^\pi(s, a) = \max_\pi \left\{ (1 - \gamma)r(s, a) + \gamma E_{s' \sim P(\cdot | s, a)} [V^\pi(s') \mid s_0 = s, a_0 = a] \right\}
\]

\[
= (1 - \gamma)r(s, a) + \gamma E_{s' \sim P(\cdot | s, a)} \max_\pi V^\pi(s')
\]

\[
= (1 - \gamma)r(s, a) + \gamma E_{s' \sim P(\cdot | s, a)} \max_\pi Q^\pi(s', \pi(s'))
\]

\[
= (1 - \gamma)r(s, a) + \gamma E_{s' \sim P(\cdot | s, a)} \max_\pi Q^\pi(s', a')
\]

\[
= (1 - \gamma)r(s, a) + \gamma E_{s' \sim P(\cdot | s, a)} \max_{a'} Q^*(s', a').
\]

Thus $Q^*$ satisfies the Bellman optimality equations.

For the converse, suppose $Q = TQ$ for some $Q$. For $\pi = \pi_Q$, this implies that $Q = (1 - \gamma)r + \gamma P\pi Q$, and so $Q = Q^\pi$, i.e. $Q$ is the action value of the policy $\pi_Q$. Now observe for any other policy $\pi'$:

\[
Q^{\pi'} - Q = (I - \gamma P^{\pi'})^{-1} r - (I - \gamma P^{\pi})^{-1} r
\]

\[
= (I - \gamma P^{\pi'})^{-1} ((I - \gamma P^{\pi}) - (I - \gamma P^{\pi'})) Q^\pi
\]

\[
= \gamma (I - \gamma P^{\pi'})^{-1} (P^{\pi'} - P^{\pi}) Q^\pi.
\]
The proof is completed by noting that $P^\pi' - P^\pi)Q^\pi \leq 0$. To see this, observe that:

$$[P^\pi' - P^\pi)Q^\pi]_{s,a} = E_{s' \sim P(\cdot|s,a)}[Q^\pi(s, \pi'(s)) - Q^\pi(s, \pi(s))] \leq 0$$

where we use the that $\pi = \pi_Q$ in the last step.

## 1.2 Planning in MDPs

Planning refers to the problem of computing $\pi^*_M$ given the MDP specification $M = (S, A, P, r, \gamma)$. This section reviews classical planning algorithms that compute $Q^*$. 

### 1.2.1 $Q$-Value Iteration

A simple algorithm is to iteratively applying the fixed point mapping: starting at some $Q$, we iteratively apply $T$:

$$Q \leftarrow T Q,$$

This is algorithm is referred to as $Q$-value iteration.

**Lemma 1.** (contraction) For any two vectors $Q, Q' \in \mathbb{R}^{|S||A|}$,

$$\|T Q - T Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

**Proof:** First, let us show that for all $s$, $\|V^Q(s) - V^{Q'}(s)\| \leq \max_{a' \in A} |Q(s, a) - Q'(s, a)|$. Assume $V^Q(s) > V^{Q'}(s)$ (the other direction is symmetric), and let $a$ be the greedy action for $Q$ at $s$. Then

$$|V^Q(s) - V^{Q'}(s)| = Q(s, a) - \max_{a' \in A} Q'(s, a') \leq Q(s, a) - Q'(s, a) \leq \max_{a \in A} |Q(s, a) - Q'(s, a)|.$$

Using this,

$$\|T Q - T Q'\|_\infty = \gamma \|PV^Q - PV^{Q'}\|_\infty = \gamma \|P(V^Q - V^{Q'})\|_\infty \leq \gamma \|V^Q - V^{Q'}\|_\infty = \gamma \max_s |V^Q(s) - V^{Q'}(s)| \leq \gamma \max_s \max_{a \in A} |Q(s, a) - Q'(s, a)| = \gamma \|Q - Q'\|_\infty$$

where the first inequality uses that each element of $P(V^Q - V^{Q'})$ is a convex average of $V^Q - V^{Q'}$, and the second inequality uses our claim above.

The following result bounds the suboptimality of the greedy policy itself, based on the error in $Q$-value function.

**Lemma 2.** [Singh and Yee [1994]] For any vector $Q \in \mathbb{R}^{|S||A|}$,

$$V^\pi_Q \geq V^* - \frac{2\|Q - Q^*\|_\infty}{1 - \gamma} \mathbf{1}.$$  

where $\mathbf{1}$ denotes the vector of all ones.
Proof: Fix state \( s \) and let \( a = \pi_Q(s) \). We have:

\[
V^*(s) - V^{\pi^*}(s) = Q^*(s, \pi^*(s)) - Q^{\pi^*}(s, a)
\]

\[
= Q^*(s, \pi^*(s)) - Q^*(s, a) + Q^*(s, a) - Q^{\pi^*}(s, a)
\]

\[
= Q^*(s, \pi^*(s)) - Q^*(s, a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[V^*(s') - V^{\pi^*}(s')]
\]

\[
\leq Q^*(s, \pi^*(s)) - Q(s, \pi^*(s)) + Q(s, a) - Q^*(s, a)
\]

\[
+ \gamma \mathbb{E}_{s' \sim P(s,a)}[V^*(s') - V^{\pi^*}(s')]
\]

\[
\leq 2\|Q - Q^*\|_\infty + \gamma \|V^* - V^{\pi^*}\|_\infty.
\]

where the first inequality uses \( Q(s, \pi^*(s)) \leq Q(s, \pi_Q(s)) = Q(s, a) \) due to the definition of \( \pi_Q \).

\[
\text{Theorem 2. (Q-value iteration convergence). Set } Q^{(0)} = 0. \text{ For } k = 0, 1, \ldots, \text{ suppose:}
\]

\[
Q^{(k+1)} = TQ^{(k)}
\]

Let \( \pi^{(k)} = \pi_Q^{(k)} \). For \( k \geq \log \frac{2}{\epsilon (1-\gamma)} / (1-\gamma) \),

\[
V^{\pi^{(k)}} \geq V^* - \epsilon I.
\]

Proof: Since \( \|Q^*\|_\infty \leq 1 \), \( Q^{(k)} = T^k Q^{(0)} \) and \( Q^* = TQ^* \), Lemma[1] gives

\[
\|Q^{(k)} - Q^*\|_\infty = \|T^k Q^{(0)} - T^k Q^*\|_\infty \leq \gamma^k \|Q^{(0)} - Q^*\|_\infty = (1 - (1 - \gamma))^k \|Q^*\|_\infty \leq \exp(-(1 - \gamma)k).
\]

The proof is completed with our choice of \( \gamma \) and using Lemma[2].

1.2.2 Policy Iteration

The policy iteration algorithm starts from an arbitrary policy \( \pi_0 \), and repeat the following iterative procedure: for \( k = 0, 1, 2, \ldots \)

1. Policy evaluation. Compute \( Q^{\pi_k} \)
2. Policy improvement. Update the policy:

\[
\pi_{k+1} = \pi Q^{\pi_k}
\]

In each iteration, we compute the Q-value function of \( \pi_k \), using the analytical form given in Equation[1.5] and update the policy to be greedy with respect to this new Q-value. The first step is often called policy evaluation, and the second step is often called policy improvement.

Lemma 3. We have that:

1. \( Q^{\pi_{k+1}} \geq TQ^{\pi_k} \geq Q^{\pi_k} \)
2. \( \|Q^{\pi_{k+1}} - Q^*\|_\infty \leq \gamma \|Q^{\pi_k} - Q^*\|_\infty \)

Proof: We start with the first part. Note that the policies produced in policy iteration are always deterministic, so \( V^{\pi_k}(s) = Q^{\pi_k}(s, \pi_k(s)) \) for all iterations \( k \) and states \( s \). Hence,

\[
TQ^{\pi_k}(s, a) = (1 - \gamma) r(s, a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q^{\pi_k}(s', a')
\]

\[
\geq (1 - \gamma) r(s, a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q^{\pi_k}(s', \pi_k(s'))
\]

\[
= Q^{\pi_k}(s, a).
\]
Using this,

\[ Q^{\pi_{k+1}}(s, a) = (1 - \gamma) r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)}[Q^{\pi_{k+1}}(s', \pi_{k+1}(s'))] \]

\[ \geq (1 - \gamma) r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)}[Q^{\pi_{k}}(s', \pi_{k+1}(s'))] \]

\[ = (1 - \gamma) r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)}[\max_{a'} Q^{\pi_{k}}(s', a')] \]

\[ = TQ^{\pi_{k}}(s, a) \]

which proves the first claim.

For the second claim,

\[ \|Q^* - Q^{\pi_{k+1}}\|_{\infty} \geq |Q^* - TQ^{\pi_{k}}|_{\infty} = |TQ^* - TQ^{\pi_{k+1}}|_{\infty} \leq \gamma \|Q^* - Q^{\pi_{k}}\|_{\infty} \]

where we have used that \( Q^* \geq Q^{\pi_{k+1}} \geq Q^{\pi_{k}} \) in second step and the contraction property of \( T(\cdot) \) (see Lemma 1 in the last step.

With this lemma, a convergence rate for the policy iteration algorithm immediately follows.

**Theorem 3. (policy iteration convergence).** Let \( \pi_0 \) be any initial policy. For \( k \geq \frac{\log \frac{1}{\epsilon}}{1 - \gamma} \), the \( k \)-th policy in policy iteration has the following performance bound:

\[ Q^{(k)} \geq Q^* - \epsilon . \]
Chapter 2

Sample Complexity with a Generative Model
2.1 The Generative Model Setting

We now characterize the optimal minimax sample complexity of estimating $Q^\star$. The results follow those in [Azar et al. 2013]. We assume that the reward function is known (and deterministic). This is often a mild assumption, particularly due to that much of the difficulty in RL is due to the uncertainty in the transition model $P$.

For this, we assume we have a access to a generative model, which can provide us with a sample $s' \sim P(\cdot | s, a)$ upon input of any state action pair. Suppose we call our simulator $N$ times at each state action pair. Let $\hat{P}$ be our empirical model, defined as follows:

$$\hat{P}(s'|s, a) = \frac{\text{count}(s', s, a)}{N}$$

where $\text{count}(s', s, a)$ is the number of times the state-action pair $(s, a)$ transitions to state $s'$. As the $N$ is the number of calls for each state action pair, the total number of calls to our generative model is $|S||A|N$.

We define $\hat{M}$ to be the empirical MDP that is identical to the original $M$, except that it uses $\hat{P}$ instead of $P$ for the transition model. When clear from context, we drop the subscript on $M$ on the values, action values, one-step variances, and variance. We let $\hat{V}^\pi, \hat{Q}^\pi, \hat{Q}^\star \hat{\pi}^\star$ denote the value function, action value function, and optimal policy in $\hat{M}$.

2.2 Sample Complexity

2.2.1 A naive approach: accurate model estimation

Note that since $P$ has a $|S|^2|A|$ parameters, a naive approach would to estimate $P$ accurately and then use our accurate model $\hat{P}$ for planning.

Theorem 4. Let $\epsilon \geq 0$. Suppose we obtain

$$\text{# samples from generative model} \geq \frac{c |S|^2|A| \log(c|S| |A|/\delta)}{(1-\gamma)^2 \epsilon^2}$$

where we sample uniformly from every state action pair. Then, with probability greater than $1 - \delta$, the following holds:

- The transition model has error bounded as:
  $$\max_{s, a} \| P(\cdot | s, a) - \hat{P}(\cdot | s, a) \|_1 \leq (1 - \gamma)^2 \epsilon / 2 .$$

- For all policies $\pi$,
  $$\| Q^\pi - \hat{Q}^\pi \|_\infty \leq \epsilon / 2$$

- The estimated $\hat{Q}^\star$ has error bounded as:
  $$\| Q^\star - \hat{Q}^\star \|_\infty \leq \epsilon$$
2.2.2 A more refined approach: using a sparse model

In the previous approach, we are able to accurately estimate the value of every policy in the unknown MDP $M$. However, with regards to planning, we only need an accurate estimate $\hat{Q}^*$ of $Q^*$, which we might hope would require less samples.

Let us start with a crude bound on the optimal action-values, which shows that an improvement is possible.

Lemma 4. (Crude Value Bounds) Let $\delta \geq 0$. With probability greater than $1 - \delta$,

$$\|Q^* - \hat{Q}^*\|_{\infty} \leq \Delta_{\delta,N}$$
$$\|Q^* - \hat{Q}^*\|_{\infty} \leq \Delta_{\delta,N},$$

where:

$$\Delta_{\delta,N} := \frac{\gamma}{1 - \gamma} \sqrt{\frac{2 \log(2|S||A|/\delta)}{N}}$$

Proof: We have:

$$\|Q^* - \hat{Q}^*\|_{\infty} = \gamma \|P^*Q^* - \hat{P}^*\hat{Q}^*\|_{\infty}$$
$$\leq \gamma \|P^*Q^* - \hat{P}^*Q^*\|_{\infty} + \gamma \|\hat{P}^*Q^* - \hat{P}^*\hat{Q}^*\|_{\infty}$$
$$= \gamma \|PV^* - \hat{P}V^*\|_{\infty} + \gamma \|\hat{P}^*(Q^* - \hat{Q}^*)\|_{\infty}$$
$$\leq \gamma \|(P - \hat{P})V^*\|_{\infty} + \gamma \|Q^* - \hat{Q}^*\|_{\infty},$$

and so we have shown that:

$$\|Q^* - \hat{Q}^*\|_{\infty} \leq \frac{\gamma}{1 - \gamma} \|(P - \hat{P})V^*\|_{\infty}$$

By applying Hoeffding’s inequality and the union bound,

$$\|(P - \hat{P})V^*\|_{\infty} = \max_{s,a} \left| E_{s' \sim P(\cdot|s,a)}[V^*(s')] - E_{s' \sim \hat{P}(\cdot|s,a)}[V^*(s')] \right| \leq \sqrt{\frac{2 \log(2|S||A|/\delta)}{N}}$$

which holds with probability greater than $1 - \delta$. This completes the proof of the first claim. The proof of the second claim is analogous.

The main result in this chapter (due to [Azar et al., 2013]) will be to improve the bound on $\hat{Q}^*$ to be optimal:

Theorem 5. (Azar et al., 2013) For $\delta \geq 0$ and with probability greater than $1 - \delta$,

$$\|Q^* - \hat{Q}^*\|_{\infty} \leq \gamma \sqrt{\frac{c \log(c|S||A|/\delta)}{1 - \gamma} \frac{1}{N}} + \frac{c \gamma}{(1 - \gamma)^2} \frac{\log(c|S||A|/\delta)}{N},$$

where $c$ is an absolute constant.

Corollary 1. Let $0 \leq \epsilon \leq \frac{1}{1 - \gamma}$. Suppose we obtain

$$\# \text{ samples from generative model} \geq \frac{c \gamma}{1 - \gamma} \frac{|S||A| \log(c|S||A|/\delta)}{\epsilon^2},$$

where we sample uniformly from every state action pair. Then, with probability greater than $1 - \delta$,

$$\|Q^* - \hat{Q}^*\|_{\infty} \leq \epsilon$$
2.2.3 Lower Bounds

Let us say that an estimation algorithm $A$, which is a map from samples to an estimate $\hat{Q}$, is $(\epsilon, \delta)$-good on MDP $M$ if $\|Q^* - \hat{Q}^*\|_\infty \leq \epsilon$ holds with probability greater than $1 - \delta$.

**Theorem 6.** (Azar et al. [2013]) There exists $\epsilon_0, \delta_0, c$ and a set of MDPs $M$ such that for $\epsilon \in (0, \epsilon_0)$ and $\delta \in (0, \delta_0)$ if algorithm $A$ is $(\epsilon, \delta)$-good on all $M \in M$, then $A$ must use a number of samples that is lower bounded as follows

$$\text{# samples from generative model} \geq \frac{c}{1 - \gamma} \frac{|S||A| \log(c|S||A|/\delta)}{\epsilon^2}.$$  

2.2.4 What about the Value of the Policy $\hat{\pi}^*$?

Ultimately, we are interested in the value $V^{\hat{\pi}^*}$ when we execute $\hat{\pi}^*$, not just an estimate $\hat{Q}^*$ of $Q^*$. The following is an immediate corollary by Lemma 2.

**Corollary 2.** For $\delta \geq 0$ and with probability greater than $1 - \delta$,

$$V^{\hat{\pi}^*} \geq V^* - \gamma \sqrt{\frac{c}{1 - \gamma} \frac{\log(c|S||A|/\delta)}{N}} - \frac{c\gamma}{(1 - \gamma)^3} \frac{\log(c|S||A|/\delta)}{N},$$

where $c$ is an absolute constant.

This bound is not sharp. Azar et al. [2013] shows that for sufficiently small $\epsilon$ — for $\epsilon \leq c'(1 - \gamma)/|S|$ (for an absolute constant $c'$) — the additional $1/(1 - \gamma)$ factor can be removed, where it becomes a lower order effect; this is an extremely stringent condition in that this amplification only becomes lower order when $\epsilon$ depends on the size of the state space. Furthermore, Sidford et al. [2018] provide a different algorithm, based on variance reduction, which removes the factor all together.

2.3 Analysis

**Lemma 5.** (Component-wise Bounds) We have that:

$$Q^* - \hat{Q}^* \leq \gamma(I - \gamma \hat{P}^{\pi^*})^{-1}(P - \hat{P})V^*$$

$$Q^* - \hat{Q}^* \geq \gamma(I - \gamma \hat{P}^{\pi^*})^{-1}(P - \hat{P})V^*$$

**Proof:** Due to the optimality of $\pi^*$ in $M$,

$$Q^* - \hat{Q}^* = Q^{\pi^*} - \hat{Q}^{\pi^*} \leq Q^{\pi^*} - \hat{Q}^{\pi^*}$$

$$= (1 - \gamma) \left( (I - \gamma P^{\pi^*})^{-1}r - (I - \gamma \hat{P}^{\pi^*})^{-1}r \right)$$

$$= (I - \gamma \hat{P}^{\pi^*})^{-1}((I - \gamma \hat{P}^{\pi^*}) - (I - \gamma P^{\pi^*}))Q^*$$

$$= \gamma(I - \gamma \hat{P}^{\pi^*})^{-1}(P^{\pi^*} - \hat{P}^{\pi^*})Q^*$$

$$= \gamma(I - \gamma \hat{P}^{\pi^*})^{-1}(P - \hat{P})V^*,$$

which proves the first claim. The second claim is left as an exercise to the reader.  

Denote the variance of any real valued \( f \) under a distribution \( D \) as:

\[
\text{Var}_D(f) := \mathbb{E}_{x \sim D}[f(x)^2] - (\mathbb{E}_{x \sim D}[f(x)])^2
\]

Slightly abusing the notation, for \( V \in \mathbb{R}^{|S|} \), we define the vector \( \text{Var}_P(V) \in \mathbb{R}^{|S||A|} \) as:

\[
\text{Var}_P(V)(s,a) := \text{Var}_{P(\cdot|s,a)}(V)
\]

Equivalently,

\[
\text{Var}_P(V) = P(V)^2 - (PV)^2.
\]

**Lemma 6.** Let \( \delta \geq 0 \). With probability greater than \( 1 - \delta \),

\[
|(P - \hat{P})V^*| \leq \sqrt{\frac{2\log(2|S||A|/\delta)}{N} \text{Var}_P(V^*) + \frac{2\log(2|S||A|/\delta)}{3N}} I.
\]

**Proof:** The claims follows from Bernstein’s inequality along with a union bound over all state-action pairs.

The key ideas in the proof are in how we bound \( \| (I - \gamma \hat{P}^\pi)^{-1} \sqrt{\text{Var}_P(V^*)} \|_\infty \) and \( \| (I - \gamma \hat{P}^\pi)^{-1} \sqrt{\text{Var}_P(V^*)} \|_\infty \).

It is helpful to define \( \Sigma_M \) as the variance of the discounted reward, i.e.

\[
\Sigma_M(s,a) := \mathbb{E}_0\left( \left( 1 - \gamma \sum_{t=0}^\infty \gamma^t r(s_t, a_t) - Q_M^\pi(s,a) \right)^2 \right)_{s_0 = s, a_0 = a}
\]

where the expectation is induced under the trajectories induced by \( \pi \) in \( M \). It is straightforward to verify that \( \| \Sigma_M \|_\infty \leq \gamma^2 \).

The following lemma shows that \( \Sigma_M \) satisfies a Bellman consistency condition.

**Lemma 7.** (Bellman consistency of \( \Sigma \)) For any MDP \( M \),

\[
\Sigma_M = \gamma^2 \text{Var}_P(V_M^\pi) + \gamma^2 P^\pi \Sigma_M
\]

where \( P \) is the transition model in MDP \( M \).

The proof is left as an exercise to the reader.

**Lemma 8.** For any policy \( \pi \) and MDP \( M \),

\[
\| (I - \gamma P^\pi)^{-1} \sqrt{\text{Var}_P(V_M^\pi)} \|_\infty \leq \sqrt{\frac{2}{1 - \gamma}}
\]

where \( P \) is the transition model in \( M \).

**Proof:** Note that \( (1 - \gamma)(I - \gamma P^\pi)^{-1} \) is matrix whose rows are a probability distribution. For a positive vector \( v \) and a distribution \( \nu \) (where \( \nu \) is vector of the same dimension of \( v \)), Jensen’s inequality implies that \( \nu \cdot \sqrt{v} \leq \sqrt{\nu \cdot v} \). This implies:

\[
\| (I - \gamma P^\pi)^{-1} \sqrt{v} \|_\infty = \frac{1}{1 - \gamma} \| (1 - \gamma)(I - \gamma P^\pi)^{-1} \sqrt{v} \|_\infty \leq \sqrt{\| \frac{1}{1 - \gamma} (I - \gamma P^\pi)^{-1} v \|_\infty} \leq \sqrt{\| \frac{2}{1 - \gamma} (I - \gamma^2 P^\pi)^{-1} v \|_\infty}.
\]
where we have used that \( \| (I - \gamma P^\pi)^{-1} v \|_\infty \leq 2 \| (I - \gamma^2 P^\pi)^{-1} v \|_\infty \) (which we will prove shortly). The proof is completed as follows: by Equation 2.1, \( \Sigma_M^\pi = \gamma^2 (I - \gamma^2 P^\pi) \text{Var}_P(V_M^\pi) \), so taking \( v = \text{Var}_P(V_M^\pi) \) and using that \( \| \Sigma_M^\pi \|_\infty \leq \gamma^2 \) completes the proof.

Finally, to see that \( \| (I - \gamma P^\pi)^{-1} v \|_\infty \leq 2 \| (I - \gamma^2 P^\pi)^{-1} v \|_\infty \), observe:

\[
\| (I - \gamma P^\pi)^{-1} \| \leq \|(1 - \gamma) I + \gamma (I - \gamma P^\pi)\| (I - \gamma^2 P^\pi)^{-1} \|_\infty \\
\leq 1 - \gamma \| (I - \gamma P^\pi)^{-1} v \|_\infty + \gamma \| (I - \gamma^2 P^\pi)^{-1} v \|_\infty \\
\leq 2 \| (I - \gamma^2 P^\pi)^{-1} v \|_\infty
\]

which proves the claim.

**Lemma 9.** Let \( \delta \geq 0 \). With probability greater than \( 1 - \delta \), we have:

\[
\text{Var}_P(V^*) \leq 2 \text{Var}_{\hat{P}}(\hat{V}^*) + \Delta_{\delta,N}' 1 \\
\text{Var}_P(V^*) \leq 2 \text{Var}_{\hat{P}}(\hat{V}^*) + \Delta_{\delta,N}' 1
\]

where

\[
\Delta_{\delta,N} := \sqrt{\frac{18 \log(6|S||A|/\delta)}{N}} + \frac{1}{(1 - \gamma)^2} \frac{4 \log(6|S||A|/\delta)}{N}.
\]

**Proof:** By definition,

\[
\text{Var}_P(V^*) = \text{Var}_P(V^*) - \text{Var}_{\hat{P}}(V^*) + \text{Var}_{\hat{P}}(V^*) \\
= P(V^*)^2 - (PV^*)^2 - \hat{P}(V^*)^2 + \hat{P}(V^*)^2 + \text{Var}_{\hat{P}}(V^*) \\
= (P - \hat{P})(V^*)^2 - (PV^*)^2 - \hat{P}(V^*)^2 + \text{Var}_{\hat{P}}(V^*)
\]

Now we bound each of these terms with Hoeffding’s inequality and the union bound. For the first term, with probability greater than \( 1 - \delta \),

\[
\| (P - \hat{P})(V^*)^2 \|_\infty \leq \sqrt{\frac{2 \log(2|S||A|/\delta)}{N}}.
\]

For the second term, again with probability greater than \( 1 - \delta \),

\[
\| (PV^*)^2 - \| PV^* \|_\infty \| PV^* - \hat{P}(V^*)^2 \| \leq 2 \| (P - \hat{P})V^* \|_\infty \leq 2 \sqrt{\frac{2 \log(2|S||A|/\delta)}{N}}.
\]

where we have used that \((\cdot)^2\) is a component-wise operation in the second step. For the last term:

\[
\text{Var}_{\hat{P}}(V^*) = \cdots 
\leq 2 \text{Var}_{\hat{P}}(V^* - \hat{V}^*) + 2 \text{Var}_{\hat{P}}(\hat{V}^*) \\
= 2 \| V^* - \hat{V}^* \|_\infty^2 + 2 \text{Var}_{\hat{P}}(\hat{V}^*) \\
= 2 \Delta_{\delta,N}^2 + 2 \text{Var}_{\hat{P}}(\hat{V}^*)
\]

To obtain a cumulative probability of error less than \( \delta \), we replace \( \delta \) in the above claims with \( \delta / 3 \). Combining these bounds completes the proof of the first claim. The above argument also shows \( \text{Var}_{\hat{P}}(V^*) \leq 2 \Delta_{\delta,N}^2 + 2 \text{Var}_{\hat{P}}(\hat{V}^*) \) which proves the second claim.

Using Lemma 6 and 9, we have the following corollary.
Corollary 3. Let $\delta \geq 0$. With probability greater than $1 - \delta$, we have:

$$(P - \hat{P}) V^* \leq c \sqrt{\frac{\text{Var}_{\hat{P}}(\hat{V}^\pi) \log(c|S||A|/\delta)}{N}} \Delta'_{\delta,N} + \Delta''_{\delta,N}$$

where

$$\Delta''_{\delta,N} := c \left( \frac{\log(c|S||A|/\delta)}{N} \right)^{3/4} + \frac{c}{1 - \gamma} \frac{\log(c|S||A|/\delta)}{N},$$

and where $c$ is an absolute constant.

2.3.1 Completing the proof

Proof:(of Theorem 5) The proof consists of bounding the terms in Lemma 8. We have:

$$\gamma \frac{\| (I - \gamma \hat{P} \pi^*)^{-1} (P - \hat{P}) V^* \|_{\infty}}{N} \leq c \sqrt{\frac{\log(c|S||A|/\delta)}{N}} \| (I - \gamma \hat{P} \pi^*)^{-1} \|_{\infty} + c \gamma \frac{\log(c|S||A|/\delta)}{N}$$

$$\leq c \gamma \sqrt{\frac{\log(c|S||A|/\delta)}{N}} + \frac{c \gamma}{1 - \gamma} \frac{\log(c|S||A|/\delta)}{N}$$

where the first step uses Corollary 3, the second uses Lemma 8, and the last step uses that $2ab \leq a^2 + b^2$ (and choosing $a, b$ appropriately). The proof of the lower bound is analogous. Taking a different absolute constant completes the proof. \[ \blacksquare \]
Chapter 3

Strategic Exploration in RL
In this lecture we will see how an agent acting in an MDP can learn a near-optimal behavior policy over time. Compared with the setting of the previous lecture on a generative model, we no longer have easy access to transitions at each state, but only have the ability to execute trajectories in the MDP. The main complexity this adds to the learning process is that the agent has to engage in exploration, that is, plan to reach new states where enough samples have not been seen yet, so that optimal behavior in those states can be learned.

The content of this chapter will be based on Brafman and Tennenholtz [2003]. In particular, we will present a version of the R-MAX algorithm, but adapted to the discounted case which we will denote as R-MAX-γ. The pseudocode of the algorithm is given in Algorithm 1. It relies on the idea of optimism in the face of uncertainty, which is common to several exploration algorithms in reinforcement learning. In a nutshell, we presume that every unknown alternative will lead to a high reward, unless we learn otherwise.

Algorithm 1 R-MAX-γ algorithm for sample efficient reinforcement learning in discounted MDPs

Input: Parameter $m$ to set known states. Accuracy parameter $\epsilon > 0$.

1. Initialize the set of known states $K = \emptyset$, counters $n(s,a) = n(s,a,s') = 0$ for all $s,a,s' \in S \times A \times S$ and $R(s,a) = 0$. Define $H = \log \frac{1}{\epsilon}/(1-\gamma)$ to the effective horizon.
2. for all episodes $i = 1,2,3,\ldots$ do
3. Let $\hat{M}$ have $\hat{P}(s'|s,a) = n(s,a,s')/n(s,a)$ and $\hat{r}(s,a) = R(s,a)/n(s,a)$.
4. Let $\hat{M}_K$ be the induced MDP (see Definition 1) and $\pi_i = \pi^*(\hat{M}_K)$ be the optimal policy in $\hat{M}_K$.
5. for $t = 0, 2, \ldots, H - 1$ do
6. Observe state $s_t$
7. If $s_t \in K$, choose $a_t = \pi_i(s_t)$, else $a_t = \arg \min_{a \in A} n(s,a)$
8. Receive reward $r_t$
9. If $s_t \notin K$, update $n(s_t, a_t) + = 1$ and $R(s_t, a_t) + = r_t$
10. If $s_{t-1} \notin K$, update $n(s_{t-1}, a_{t-1}, s_t) + = 1$
11. end for
12. If a state becomes known, i.e. $n(s,a) \geq m$ for all $a \in A$, update $K = K \cup \{s\}$.
13. end for

Concretely, the algorithm maintains an estimate of the transition probabilities $P(s'|s,a)$ for all the neighbors $s'$ of a state $s$, given an action $a$. It also estimates the reward $r(s,a)$. Once the algorithm has visited $s$ adequately often to ensure that these estimates are all accurate, it declares the state as known. Learning is complete when all the states are known. During a run of the algorithm, whenever it is in a known state, it already knows the optimal action to take and follows this action. However, when the algorithm is in an unknown state, it explores by picking the action chosen least often in the state so far.

While we will mostly focus on the statistical properties of the algorithm, the computational aspects are relatively straightforward. Within an episode, the main computational burden is in the computation of an optimal policy for the induced MDP $M_K$ in line 3. This can be done, for example, using the value iteration algorithm from Chapter 1.2 as the MDP and reward function are fully known in this step. Since reasoning over the infinite horizon can be tricky computationally, a common trick is to restrict the step in Algorithm 1 to computing a non-stationary $H$-step optimal policy instead. That is, we find the policy which maximizes the expected discounted reward over just $H$ time steps. Such a policy can be easily computed via dynamic programming, but is required to be non-stationary. That is, it might choose different actions for the same state visited at different values of $t$. Our choice of $H$ ensures that the
infinite-horizon value functions and $H$-step value functions are at most $\epsilon$ apart, so that none of the subsequent theory is affected.

In order to more concretely discuss the algorithm, we need some important definitions. Given an MDP $M$ and a set $K$ of known states, we next define the notion of an Induced MDP.

**Definition 1 (Induced MDP).** Let $M$ be an MDP parametrized by $(S, A, P, r, \gamma)$ with $K \subseteq S$ being a subset of states. Based on this set, we define the induced MDP $M_K$ parametrized by $(S, A, P_K, r_K, \gamma)$ in the following manner. For each $s \in K$, we define

$$P_{M_K}(s'|s, a) = P_M(s'|s, a) \quad \text{and} \quad r_{M_K}(s, a) = r_M(s, a).$$

For all the $s \notin K$, we define

$$P_{M_K}(s'|s, a) = 1(s' = s) \quad \text{and} \quad r_{M_K}(z|s, a) = 1(z = 1).$$

Thus, an induced MDP given a set $K$ of known states is an optimistic process where we receive a reward of 1 (recall that the rewards $r_i \in [0, 1]$ so that 1 is the largest attainable reward) no matter which action we try in an unknown state. Furthermore, once we enter such an unknown state, we stay there and keep collecting this reward for the remainder of the episode. On the known states, naturally the transition and reward distributions follow their known behavior.

We will now analyze the R-MAX-$\gamma$ algorithm, and provide a bound on the number of episodes before which it finds an $\epsilon$-optimal policy. We will prove the following theorem.

**Theorem 7.** Let $s_{i,t}$ be the state visited by the R-MAX-$\gamma$ algorithm at round $t$ in episode $i$. For any $0 \leq \epsilon, \delta < 1$, with probability at least $1 - \delta$, $V^*_M(s_{i,t}) \geq V^*_M(s_{i,t}) - \epsilon$, for all but $O \left( \frac{H^3S^2A \log \frac{S^2A}{\delta}}{\epsilon^2} \right)$ actions in the MDP.

In words, the algorithm finds policies such that those policies induce near optimal value functions for all but a bounded number of rounds. Note that this does not imply that the algorithm behaves sub-optimally for the first $O \left( \frac{H^3S^2A \log \frac{S^2A}{\delta}}{\epsilon^2} \right)$ rounds only. The dynamics of the MDP might be such that an unknown state is encountered with a small chance only, in which case the algorithm learns whenever it encounters these states. The guarantee also does not preclude settings where after some initial exploration, the algorithm encounters a state with a small value from which escape is not possible under the dynamics. In such a case, the guarantee of the algorithm trivially holds as any policy is optimal in that state.

An alternative optimality condition we might desire from the algorithm is that it finds a near-optimal policy, that is a policy whose expected reward is within $\epsilon$ of the optimal, when taking expectations over the start state as well. This is not the guarantee provided here, and in general requires some mixing conditions on the MDP which we do not consider here. A common way to ensure such mixing in practice is by assuming the ability to reset to the initial state distribution during the training of an agent. When such a reset ability is available, the guarantee provided here can be further strengthened into approximate optimality of the policy.

In order to prove the result, we will introduce a number of key concepts in understanding exploration in reinforcement learning. Throughout the analysis, we will abuse our notation $r$ to also refer to the expected reward, given a (state, action) pair.

We start with a basic result which was first introduced in [Kearns and Singh 2002] under the name of a simulation lemma.

**Lemma 10 (Simulation lemma for MDPs).** Let $M$ and $M'$ be two MDPs with the same state and action spaces. If the transition and reward functions of these MDPs satisfy

$$\sum_{s' \in S} |P_M(s'|s, a) - P_{M'}(s'|s, a)| \leq \epsilon_1, \quad \forall s \in S \text{ and } a \in A,$$

and

$$|r_M(s, a) - r_{M'}(s, a)| \leq \epsilon_2, \quad \forall s \in S \text{ and } a \in A.$$
Then for every stationary policy $\pi$, the two MDPs satisfy \( \|V^\pi_M - V^\pi_{M'}\|_\infty \leq \gamma_1 \epsilon_1 + \epsilon_2 \).

The lemma is called a simulation lemma as it tells how much error we incur in evaluating policies if we build an approximate simulator $M'$ for the true process $M$.

**Proof:** The lemma is proved using the conditions on the transition and reward distributions, along with the Bellman equations for value functions \([1, 2]\). For any state $s$, we have
\[
|V^\pi_M(s) - V^\pi_{M'}(s)| \leq (1 - \gamma)\epsilon_2 + \gamma \sum_{s' \in S} (P_M(s'|s, \pi(s))V^\pi_M(s') - P_{M'}(s'|s, \pi(s))V^\pi_{M'}(s'))
\]
\[
\leq (1 - \gamma)\epsilon_2 + \gamma \sum_{s' \in S} P_M(s'|s, a)(V^\pi_M(s') - V^\pi_{M'}(s')) + \gamma \sum_{s' \in S} V^\pi_{M'}(s')(P_M(s'|s, a) - P_{M'}(s'|s, a))
\]
\[
\leq (1 - \gamma)\epsilon_2 + \gamma\|V^\pi_M - V^\pi_{M'}\|_\infty + \gamma \epsilon_1.
\]

Note that here we have used the normalization of value functions, that is $0 \leq V^\pi_M(s) \leq 1$. Since the inequality holds for any state, we can take the max on the LHS and rearrange terms to complete the proof.

The next lemma really formalizes our intuition that the optimal policy in the induced MDP encourages exploration of the currently unknown states. We will show that either the best policy $\pi_i$, learned using the induced MDP at an episode $i$ is already good, or it has a high chance of taking us to an unknown state. We will use the notation
\[
\mathbb{P}^\pi_M[\text{escape from } K|s_0 = s] := \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi_M(s_t \notin K|s_0, \ldots, s_{t-1} \in K).
\]

That is, $\mathbb{P}^\pi_M[\text{escape from } K|s_0 = s]$ is the discounted probability of reaching an unknown state when executing $\pi$ in the original MDP $M$, starting from the state $s$.

**Lemma 11** (Induced inequalities). Let $M$ be an MDP with $K$ being the set of known states. Let $M_K$ be the induced MDP (Definition\([1]\)) with respect to $K$ and $M$. For any stationary policy $\pi$ and state $s \in S$ we have
\[
V^\pi_{M_K}(s) \geq V^\pi_M(s) \quad \text{and} \quad V^\pi_M(s) \geq V^\pi_{M_K}(s) - \mathbb{P}^\pi_M[\text{escape from } K|s_0 = s].
\]

The lemma has two implications. First it formalizes the notion that the induced MDP $M_K$ is indeed an optimistic version of $M$, since it ascribes higher values to each state under every policy. At the same time, the optimism is not uncontrolled. The values ascribed by $M_K$ to a policy $\pi$ are higher only if the policy has a substantial probability of visiting an unknown state, and hence is useful for exploration.

**Proof:** The first inequality is a direct consequence of the definition of $M_K$. If $s \notin K$, it is immediate since we get the maximum reward of 1 at each time-step, while never leaving this state. If $s \in K$, then our immediate reward is identical to that in $M$. At the next step, we either stay in $K$, or leave. If we leave then we will obtain the largest reward for the remaining time steps. If we stay, we obtain the same reward as that in $M$. Thus we never obtain a smaller reward in $M_K$ by definition.

For the second part, we have
\[
|V^\pi_M(s) - V^\pi_{M_K}(s)| \leq 1(s \notin K) + 1(s \in K)\gamma \sum_{s' \in S} P_M(s'|s, \pi(s))V^\pi_M(s') - P_{M_K}(s'|s, \pi(s))V^\pi_{M_K}(s')
\]
\[
= 1(s \notin K) + 1(s \in K)\gamma \sum_{s' \in S} P_M(s'|s, \pi(s))(V^\pi_M(s') - V^\pi_{M_K}(s'))
\]
\[
\leq 1(s \notin K) + 1(s \in K)\gamma P_M(s' \notin K|s, \pi(s)) + 1(s \in K)\gamma \sum_{s' \in K} P_M(s'|s, \pi(s))(V^\pi_M(s') - V^\pi_{M_K}(s')).
\]
Here the first inequality holds since the two value functions can differ by at most the maximum value of 1 if the starting state is unknown. The following equality holds as the transition models under $M$ and $M_K$ are identical when $s \in K$. Now unrolling the summation in the last inequality further yields the statement of the lemma.

Given the lemma, we have a particularly useful corollary. It says that the policy $\pi_i$ computed in each episode of Algorithm 1 is near optimal, with the error being the probability of leaving the known state set.

**Corollary 4 (Implicit Explore-Exploit).**

$$V_M^{\pi^*(M_K)}(s) \geq V_M^*(s) - \mathbb{P}_M^{\pi^*(M_K)}[\text{escape from } K|s_0 = s]$$

**Proof:** By the lemma, we have

$$V_M^{\pi^*(M_K)}(s) \geq V_M^{\pi^*(M_K)}(s) - \mathbb{P}_M^{\pi^*(M_K)}[\text{escape from } K|s_0 = s]$$

$$\geq V_M^{\pi^*(M)}(s) - \mathbb{P}_M^{\pi^*(M_K)}[\text{escape from } K|s_0 = s]$$

$$\geq V_M^{\pi^*(M)}(s) - \mathbb{P}_M^{\pi^*(M_K)}[\text{escape from } K|s_0 = s].$$

Here the first inequality follows from Lemma 11 applied with $\pi = \pi^*(M_K)$, second inequality uses that $\pi^*(M_K)$ is the optimal policy in $M_K$ and hence obtains a higher reward than $\pi^*(M)$ and the third inequality follows from the optimism of $M_K$ shown in Lemma 11.

**Proof of Theorem 7.**

We now have most of the ingredients for the theorem. In order to prove the theorem, we need to ensure that $m$ is large enough that when a state is declared known, then its transition and reward functions are reasonably accurate. For now, let us assume that $m$ is chosen large enough so that the induced approximate MDP $\widehat{M}_K$ is a good approximation to the true induced MDP $M_K$. Based on Lemma 10, we will assume that $m$ is large enough so that the value functions of these two MDPs are at most $\epsilon/2$ different in any state. Then, applying this closeness twice, we see that

$$V_{\widehat{M}_K}^{\pi_i}(s) \geq V_{\widehat{M}_K}^{\widehat{\pi_i}}(s) - \frac{\epsilon}{2} \geq V_{\widehat{M}_K}^{\pi_i}(s) - \frac{\epsilon}{2} \geq V_{M_K}^{\pi^*}(s) - \epsilon.$$

Combining with Lemma 11 we see that for any starting state we have

$$V_{\widehat{M}_K}^{\pi_i}(s) \geq V_{M_K}^{\pi^*}(s) - \epsilon - \mathbb{P}_{\widehat{M}_K}^{\pi_i}[\text{escape from } K|s_0 = s]$$

$$\geq V_{M}^{\pi^*}(s) - \epsilon - \mathbb{P}_{\widehat{M}_K}^{\pi_i}[\text{escape from } K|s_0 = s],$$

where the first inequality is combining Lemma 11 with with the earlier bound, and the second inequality uses the optimism of the induced MDP $M_K$.

Thus, either the policy $\widehat{\pi}_i$ is at most $2\epsilon$ suboptimal, or it visits an unknown state with probability at least $\epsilon$. Intuitively, this means that we visit an unknown state at least every $1/\epsilon$ steps, if $\widehat{\pi}_i$ is not already near optimal. The total number of visits to unknown states are bounded by $mSA$. This is because for each unknown state $s$, we need to try every action $a$ at least $m$ times before $s$ becomes known. Since we try the least frequently chosen action each time, it is ensured that each action is chosen exactly $m$ times before $s$ becomes known. Consequently, we need $O(mSA/\epsilon)$ episodes in order to ensure that every state is known and the algorithm can certifiably have a near optimal policy. In other words, with $O(mSAH/\epsilon)$ actions in the MDP, the agent is guaranteed to have marked all the states as known.
In order to obtain the theorem statement, we set $m = O\left(\frac{8H^2}{\epsilon^2} \log \frac{SA}{\delta}\right)$. This number is based on satisfying the condition of Lemma 10 with $\gamma \epsilon_1 / (1 - \gamma) = \epsilon/2$, along with Lemma 8.5.6 of Kakade [2003]. Plugging this value of $m$ in our bound on the number of samples as a function of $m$ above completes the proof of the theorem.
Bibliography


Appendix A

3.1 Concentration

Lemma 12. (Hoeffding’s inequality) Suppose $X_1, X_2, \ldots, X_n$ are a sequence of independent, identically distributed (i.i.d.) random variables with mean $\mu$. Let $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Suppose that $X_i \in [b_-, b_+]$ with probability 1, then

$$P(\bar{X}_n \geq \mu + \epsilon) \leq e^{-2n\epsilon^2/(b_+ - b_-)^2}.$$ 

Similarly,

$$P(\bar{X}_n \leq \mu - \epsilon) \leq e^{-2n\epsilon^2/(b_+ - b_-)^2}.$$ 

The Chernoff bound implies that with probability $1 - \delta$:

$$\bar{X}_n - EX \leq (b_+ - b_-)\sqrt{\ln(1/\delta)/(2n)}.$$ 

Lemma 13. (Bernstein’s inequality) Suppose $X_1, \ldots, X_n$ are independent random variables. Let $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$, $\mu = \mathbb{E}\bar{X}_n$, and $\text{Var}(X_i)$ denote the variance of $X_i$. If $X_i - EX_i \leq b$ for all $i$, then

$$P(\bar{X}_n \geq \mu + \epsilon) \leq \exp \left[ -\frac{n^2\epsilon^2}{2 \sum_{i=1}^{n} \text{Var}(X_i) + 2nb\epsilon/3} \right].$$ 

If all the variances are equal, the Bernstein inequality implies that, with probability at least $1 - \delta$,

$$\bar{X}_n - EX \leq \sqrt{2\text{Var}(X) \ln(1/\delta)/n} + \frac{2b\ln(1/\delta)}{3n}.$$