Chapter 0

Notation

0.1 Notation

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

- For a vector \( v \), we let \((v)^2\), \(\sqrt{v}\), and \(|v|\) be the component-wise square, square root, and absolute value operations.
- Inequalities between vectors are elementwise, e.g. for vectors \( v, v' \), we way \( v \leq v' \), if the inequality holds elementwise.
- For a vector \( v \), we refer to the \( j \)-th component of this vector by either \( v(j) \) or \([v]_j\).
- Denote the variance of any real valued \( f \) under a distribution \( D \) as:
  \[
  \text{Var}_D(f) := E_{x \sim D}[f(x)^2] - (E_{x \sim D}[f(x)])^2
  \]
- 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:
  \[
  TQ := (1 - \gamma)r + \gamma PV_Q.
  \]
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 $s$.
  We use $P_{s,a}$ to denote the vector $P(\cdot | s, a)$.
- **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.

### 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 $M = (S, A, P, r, \gamma, \mu)$.

### 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. $\pi: H \rightarrow A$ where $H$ 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
Lecture 1:

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:

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

(1.1)

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 \( s \in \mathcal{S}, a \in \mathcal{A} \),

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

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

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

It is helpful to view \( V_\pi \) as vector of length \( S \) and \( Q_\pi \) 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 \( P(s', a|s, a) \) 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 \( P_\pi (s, a), (s', a') = P(s'|s, a)p_i(a|s') \). With this notation, it is straightforward to verify:

\[
Q_\pi = (1 - \gamma)r + \gamma PV_\pi \\
Q_\pi = (1 - \gamma)r + \gamma P_\pi Q_\pi
\]

where \( I \) is the identity matrix.

The above implies that:

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

To see that the \((I_{|S|} - \gamma P_\pi)\) 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 \\
\geq \| x \|_\infty - \gamma \| P_\pi x \|_\infty \quad \text{(triangular inequality for norms)} \\
\geq \| x \|_\infty - \gamma \| x \|_\infty \quad \text{(each element of } P_\pi x \text{ is a convex average of } x) \\
= (1 - \gamma)\| x \|_\infty > 0 \quad \text{\( (\gamma < 1, x \neq 0) \)}
\]

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

### 1.14 Bellman optimality equations

Due to the Markov structure, there exists a single stationary and deterministic policy that simultaneously maximizes \( V_\pi (s) \) for all \( s \in S \) and maximizes \( Q_\pi (s, a) \) for all \( s \in S, a \in A \) \cite{Puterman1994}; 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 \cite{Bellman1956}: for all \( s \in S, a \in A \),

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

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

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

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 \( Q \in \mathbb{R}^{|S\times A|} \) into a vector of length \( |S| \).

\[
V_Q(s) := \max_{a \in A} Q(s, a).
\]
The *Bellman optimality operator* $\mathcal{T}_M : \mathbb{R}^{|S \times A|} \to \mathbb{R}^{|S \times A|}$ is defined as follows: when applied to some vector $Q \in \mathbb{R}^{|S \times A|}$,

$$\mathcal{T}Q := (1 - \gamma)r + \gamma PV_Q.$$  \hfill (1.7)

This allows us to rewrite Equation 1.6 in the concise form: $Q^* = \mathcal{T}Q^*$, i.e. $Q^*$ is a fixed point of the operator $\mathcal{T}$. The classic result of [Bellman, 1956] shows that if $Q$ satisfies $Q = \mathcal{T}Q$, 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 $Q = \mathcal{T}Q$.

**Proof:** First observe that:

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

$$= (1 - \gamma)\mathbb{E} \left[ r(s_0, a_0) + \ldots + r(s_{\tau-1}, a_{\tau-1}) + \gamma^\tau E \left[ \sum_{t=0}^{\infty} \gamma^t r(s_{t+\tau}, a_{t+\tau}) | s_\tau = s, a_\tau = a \right] \big| s_0 = s, a_0 = a \right]$$

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

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 $Q^* = \mathcal{T}Q^*$. We need only consider deterministic policies. We have:

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

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

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

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

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

For the converse, suppose $Q = \mathcal{T}Q$ for some $Q$. For $\pi = \pi_Q$, this implies that $Q = (1 - \gamma)r + PQ$, and so $Q = Q^*$, 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:

\[\left[P^\pi' - P^\pi\right]_{s,a} = \mathbb{E}_{s' \sim P(s,a)}[Q^\pi(s, \pi'(s)) - Q^\pi(s, \pi(s))] \leq 0\]

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

\[\boxed{}\]

### 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 \(T:\)

\[Q \leftarrow TQ,\]

This is algorithm is referred to as \(Q\)-value iteration.

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

\[\|TQ - TQ'\|_\infty \leq \gamma \|Q - Q'\|_\infty\]

**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,

\[\|TQ - TQ'\|_\infty = \gamma \|PV_Q - PV_{Q'}\|_\infty = \gamma \|P(V_Q - V_{Q'})\|_\infty \leq \gamma \|V_Q - V_{Q'}\|_\infty \leq \gamma \max_s |V_Q(s) - V_{Q'}(s)| \leq \gamma \max_s \max_{a \in A} |Q(s, a) - Q'(s, a)| = \gamma \|Q - Q'\|_\infty\]

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.** \([\text{Singh and Yee, 1994}]\) For any vector \(Q \in \mathbb{R}^{|S||A|}\),

\[V^{\pi_Q} \geq V^* - \frac{2\|Q - Q^*\|_\infty}{1 - \gamma}\]

where \(\mathbf{1}\) denotes the vector of all ones.
Proof: Let \( a = \pi_Q(s) \). We have:

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

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

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

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

\[
+ \gamma E_{s' \sim P(s, a)}[V^*(s') - V^{\pi_Q}(s')]
\]

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

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

\[\square\]

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

\[\square\]

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. \( Q^{\pi_{k+1}} \geq T Q^{\pi_k} \geq Q^{\pi_k} \)

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,

\[
T Q^{\pi_k}(s, a) = (1 - \gamma)r(s, a) + \gamma E_{s' \sim P(s, a)}[Q^{\pi_k}(s', \max_{a'} Q^{\pi_k}(s', a'))]
\]

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

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

\[ \|Q^\ast - Q^{\pi_k+1}\|_\infty \geq \|Q^\ast - TQ^{\pi_k}\|_\infty = \|TQ^\ast - TQ^{\pi_k+1}\|_\infty \leq \gamma = \|Q^\ast - Q^{\pi_k}\|_\infty \]

where we have used that \( Q^\ast \geq Q^{\pi_k+1} \geq Q^{\pi_k} \) 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 \( k \geq \log \frac{1}{\epsilon/(1 - \gamma)} \), the \( k \)-th policy in policy iteration has the following performance bound:

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

Sample Complexity with a Generative Model
2.1 The Generative Model Setting

We now characterize the optimal minimax sample complexity of estimating $Q^*$. The results follow those in [Azar et al., 2013].

Assume we have 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 $\hat{V}^\pi, \hat{Q}^\pi, \hat{Q}^* \pi^*$ 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

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

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 $\hat{Q}^*$ has error bounded as:
  $$\| Q^* - \hat{Q}^* \|_\infty \leq \epsilon$$
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 \( 1 - \delta \),

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

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,

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

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 \( 1 - \delta \),

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

where \( c \) is an absolute constant.

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

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

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

To be added...

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 \( M \),

\[
Q^* - \hat{Q}^* = Q^{\pi^*} - \hat{Q}^{\pi^*} \\
\leq Q^{\pi^*} - \hat{Q}^{\pi^*} \\
\leq (I - \gamma P^{\pi^*})^{-1} r - (I - \gamma \hat{P}^{\pi^*})^{-1} r \\
\leq (I - \gamma \hat{P}^{\pi^*})^{-1} ((I - \gamma \hat{P}^{\pi^*}) - (I - \gamma P^{\pi^*})) Q^* \\
\leq \gamma (I - \gamma \hat{P}^{\pi^*})^{-1} (P^{\pi^*} - \hat{P}^{\pi^*}) Q^* \\
\leq \gamma (I - \gamma \hat{P}^{\pi^*})^{-1} (P - \hat{P}) V^*,
\]

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:

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

With respect to policy \( \pi \), define the “one-step” variance in the MDP \( M \) as:

\[
\sigma_\pi^2 (s,a) := \text{Var}_{P(\cdot | s,a)}(V^\pi_M),
\]

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

\[
\sigma_\pi^2 = P(V^\pi_M)^2 - (PV^\pi_M)^2.
\]

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

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

**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 \( \| (I - \gamma \hat{P}^{\pi^*})^{-1} \sqrt{\sigma^2} \|_\infty \) and \( \| (I - \gamma \hat{P}^{\pi^*})^{-1} \sqrt{\sigma^2} \|_\infty \).

It is helpful to define \( \Sigma_\pi^2_M \) as the variance of the discounted reward, i.e.

\[
\Sigma_\pi^2_M(s,a) := \mathbb{E} \left[ \left( 1 - \gamma \sum_{t=0}^\infty \gamma^t r(s_t, a_t) - Q^*_M(s,a) \right)^2 \right] |s_0 = s, a_0 = a|
\]

where the expectation is induced under the trajectories induced by \( \pi \) in \( M \). The following lemma shows that \( \Sigma_\pi^2_M \) satisfies a Bellman consistency condition.
Lemma 7. (*Bellman consistency of $\Sigma_M^\pi$*) For any MDP $M$,

$$
\Sigma_M^\pi = \gamma^2 \Sigma_M^\pi + \gamma^2 P^\pi \Sigma_M^\pi
$$

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 $M$,

$$
\| (I - \gamma P^\pi)^{-1} \sqrt{\sigma_M^\pi} \|^2_\infty \leq 2 \sqrt{\frac{1}{1 - \gamma}} \| \Sigma_M^\pi \|_\infty \leq 2 \sqrt{\frac{1}{1 - \gamma}}
$$

Proof: To be added....

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

$$
\sigma^\pi \leq 2\hat{\sigma}^\pi + \Delta_{\delta,N}' \mathbb{I}
$$

$$
\sigma^* \leq 2\hat{\sigma}^* + \Delta_{\delta,N}' \mathbb{I}
$$

where

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

Proof: By definition of $\sigma^\pi$ and $\hat{\sigma}^\pi$,

$$
\sigma^\pi(s, a) = \sigma^\pi(s, a) - \text{Var}_{\overline{P}_{(s,a)}}(V^*) - \text{Var}_{\overline{P}_{(s,a)}}(V^*)
$$

$$
= (P(V^*)^2)(s, a) - (PV^*)^2(s, a) - (\overline{P}(V^*)^2)(s, a) + (\overline{P}V^*)(s, a) + \text{Var}_{\overline{P}_{(s,a)}}(V^*)
$$

$$
= ((P - \overline{P})(V^*)^2)(s, a) - ((PV^*)^2 - (\overline{P}V^*)^2)(s, a) + \text{Var}_{\overline{P}_{(s,a)}}(V^*)
$$

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

$$
\|(P - \overline{P})(V^*)^2\|_\infty \leq \sqrt{2 \log(2|S||A|/\delta)}.
$$

For the second term,

$$
\|(PV^*)^2 - (\overline{P}V^*)^2\|_\infty \leq \|PV^* + \overline{P}V^*\|_\infty \|PV^* - \overline{P}V^*\|_\infty \leq 2\|(P - \overline{P})V^*\|_\infty \leq 2 \sqrt{2 \log(2|S||A|/\delta)}.
$$

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

$$
\text{Var}_{\overline{P}_{(s,a)}}(V^*)(s, a) = \text{Var}_{\overline{P}_{(s,a)}}(V^* - \overline{V}^* + \overline{V}^*)
$$

$$
\leq 2\text{Var}_{\overline{P}_{(s,a)}}(V^* - \overline{V}^*) + 2\text{Var}_{\overline{P}_{(s,a)}}(\overline{V}^*)
$$

$$
= 2\|V^* - \overline{V}^*\|^2_\infty + 2\hat{\sigma}(s, a)
$$

$$
= 2\Delta_{\delta,N}^2 + 2\hat{\sigma}(s, a).
$$

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 $\text{Var}_{\overline{P}_{(s,a)}}(V^*)(s, a) \leq 2\Delta_{\delta,N}^2 + 2\hat{\sigma}(s, a)$ which proves the second claim.
Corollary 2. Let $\delta \geq 0$. With probability greater than $1 - \delta$, we have:
\[
| (P^{*} - \tilde{P}^{*}) V^{*} | \leq c \sqrt{\frac{\sigma^{*} \log(c|S||A|/\delta)}{N}} + \Delta''_{\delta,N} \mathbb{I}
\]
\[
| (P^{*} - \tilde{P}^{*}) V^{*} | \leq c \sqrt{\frac{\sigma^{*} \log(c|S||A|/\delta)}{N}} + \Delta''_{\delta,N} \mathbb{I}
\]
where $c$ is an absolute constant and where:
\[
\Delta''_{\delta,N} := c \left( \frac{\log(c|S||A|/\delta)}{N} \right)^{3/4} + \frac{c}{1 - \gamma} \frac{\log(c|S||A|/\delta)}{N}.
\]

2.3.1 Completing the proof

Proof: (of Theorem 5) The proof consists of bounding the terms in Lemma 5. We have:
\[
\| (I - \gamma \tilde{P}^{*})^{-1} (P - \tilde{P}) V^{*} \|_{\infty} \leq \frac{c}{1 - \gamma} \sqrt{\frac{\log(c|S||A|/\delta)}{N}} + \frac{1}{1 - \gamma} \Delta''_{\delta,N}
\]
\[
\leq 2 \sqrt{\frac{1}{1 - \gamma}} \frac{\log(c|S||A|/\delta)}{N} + \frac{1}{1 - \gamma} \Delta''_{\delta,N}
\]
\[
= 2c \sqrt{\frac{1}{1 - \gamma}} \frac{\log(c|S||A|/\delta)}{N} + \frac{c}{1 - \gamma} \left( \frac{\log(c|S||A|/\delta)}{N} \right)^{3/4} + \frac{c}{(1 - \gamma)^2} \frac{\log(c|S||A|/\delta)}{N}
\]
\[
\leq 3 \sqrt{\frac{1}{1 - \gamma}} \frac{\log(c|S||A|/\delta)}{N} + 2 \frac{c}{(1 - \gamma)^2} \frac{\log(c|S||A|/\delta)}{N},
\]
where the last step uses that $ab \leq a^2 + b^2$. The proof of the lower bound is analogous. Taking a different absolute constant completes the proof. □
Bibliography


Appendix A

Appendix

2.1 Concentration

**Lemma 10.** (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 $ar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. 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 11.** (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 $1 - \delta$,

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

21