Reinforcement Learning: Theory and Algorithms

Alekh Agarwal   Nan Jiang   Sham M. Kakade

May 19, 2019

Contents

Chapter 0

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 $\mathcal{D}$ as:

$$\mathrm{Var}_{\mathcal{D}}(f) := E_{x\sim\mathcal{D}}[f(x)^2] - (E_{x\sim\mathcal{D}}[f(x)])^2$$

- It is helpful to overload notation and let P also refer to a matrix of size $(\mathcal{S} \cdot \mathcal{A}) \times \mathcal{S}$ where the entry $P_{(s,a),s'}$ is equal to $P(s'|s,a)$. We also will define P^π to be the transition matrix on state-action pairs induced by a deterministic policy π. 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,

$$\begin{aligned}
Q^\pi &= (1-\gamma)r + \gamma P V^\pi \\
Q^\pi &= (1-\gamma)r + \gamma P^\pi Q^\pi \\
Q^\pi &= (1-\gamma)(I - \gamma P^\pi)^{-1} r
\end{aligned}$$

- For a vector $Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, denote the greedy policy and value as:

$$\begin{aligned}
\pi_Q(s) &:= \mathrm{argmax}_{a\in\mathcal{A}} Q(s,a) \\
V_Q(s) &:= \max_{a\in\mathcal{A}} Q(s,a)..
\end{aligned}$$

- For a vector $Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, the *Bellman optimality operator* $\mathcal{T} : \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is defined as:

$$\mathcal{T}Q := (1-\gamma)r + \gamma P V_Q\,. \tag{0.1}$$

Chapter 1

MDP Preliminaries

Reinforcement Learning and Bandits **Spring 2019**

Markov Decision Processes

Instructors: Alekh Agarwal, Sham Kakade *Lecture 1*

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 $\mathcal{S}$. In this course we only consider finite state spaces.
- Action space $\mathcal{A}$. In this course we only consider finite action spaces.
- Transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ is the space of probability distributions over $\mathcal{S}$ (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 $P(\cdot \mid s,a)$.
- Reward function $r : \mathcal{S} \times \mathcal{A} \to [0,1]$. $r(s,a)$ 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 $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, 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 $a_t \in \mathcal{A}$, 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 $\mu \in \Delta(\mathcal{S})$. When μ is of importance to the discussion, we include it as part of the MDP definition, and write $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$.

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. $\pi : \mathcal{H} \to \mathcal{A}$ where $\mathcal{H}$ is the set of all possibly trajectories. A deterministic, *stationary* policy $\pi : \mathcal{S} \to \mathcal{A}$ 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 $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, 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 $s \in \mathcal{S}$.

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

$$V_M^\pi(s) = (1-\gamma)\mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid \pi, s_0 = s\Big].$$

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

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

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

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

$$\max_\pi V_M^\pi(s) \tag{1.1}$$

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

Example 1.1 (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 $(1-\gamma)\gamma^d$, where d is the number of steps required by the policy to reach the goal state.

Example 1.2 (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 1.3 (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^π and Q^π satisfy the following *Bellman consistency equations*: for all $s \in \mathcal{S}, a \in \mathcal{A}$,

$$\begin{aligned} V^\pi(s) &= Q^\pi(s, \pi(s)). \\ Q^\pi(s,a) &= (1-\gamma) r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\big[V^\pi(s')\big], \end{aligned} \tag{1.2}$$

where we are treating π as a deterministic policy.

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

$$Q^\pi = (1-\gamma)r + \gamma P V^\pi \tag{1.3}$$
$$Q^\pi = (1-\gamma)r + \gamma P^\pi Q^\pi \,. \tag{1.4}$$

The above implies that:

$$Q^\pi = (1-\gamma)(I - \gamma P^\pi)^{-1} r \tag{1.5}$$

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

$$\begin{aligned} \|(I-\gamma P^\pi)x\|_\infty &= \|x - \gamma P^\pi x\|_\infty \\ &\geq \|x\|_\infty - \gamma \|P^\pi x\|_\infty && \text{(triangular inequality for norms)} \\ &\geq \|x\|_\infty - \gamma \|x\|_\infty && \text{(each element of } P^\pi x \text{ is a convex average of } x\text{)} \\ &= (1-\gamma)\|x\|_\infty > 0 && (\gamma < 1,\, x \neq 0) \end{aligned}$$

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 $V^\pi(s)$ for all $s \in \mathcal{S}$ and maximizes $Q^\pi(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$ [Puterman, 1994]; we denote this *optimal policy* as $\pi^\star_M$ (or $\pi^\star$). We use $V^\star$ and $Q^\star$ as a shorthand for $V^{\pi^\star}$ and $Q^{\pi^\star}$, respectively.

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

$$\begin{aligned} V^\star(s) &= \max_{a \in \mathcal{A}} Q^\star(s,a). \\ Q^\star(s,a) &= (1-\gamma)r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\big[V^\star(s')\big]. \end{aligned} \tag{1.6}$$

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

$$\pi_Q(s) := \mathrm{argmax}_{a \in \mathcal{A}} Q(s,a) \,.$$

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

$$\pi^\star = \pi_{Q^\star} \,.$$

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

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

The *Bellman optimality operator* $\mathcal{T}_M : \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is defined as follows: when applied to some vector $Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$,

$$\mathcal{T}Q := (1-\gamma)r + \gamma P V_Q \,. \tag{1.7}$$

This allows us to rewrite Equation 1.6 in the concise form: $Q^\star = \mathcal{T}Q^\star$, i.e. $Q^\star$ 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 $Q = Q^\star$. We state the result below formally.

Theorem 1.4. *Let $Q^\star(s,a) = \max_{\pi\in\Pi} Q^\pi(s,a)$ where Π is the space of all (non-stationary and randomized) policies. We have that*

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

Proof: First observe that:

$$\begin{aligned}
& Q^\star(s,a) \\
= \; & (1-\gamma)\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) \;\big|\; s_0=s, a_0=a\Big] \\
= \; & (1-\gamma)\mathbb{E}\Big[r(s_0,a_0)+\ldots r(s_{\tau-1},a_{\tau-1}) + \gamma^\tau E[\sum_{t=\tau}^{\infty}\gamma^t r(s_{t+\tau},a_{t+\tau})|s_\tau=s,a_\tau=a] \;\big|\; s_0=s,a_0=a\Big] \\
= \; & (1-\gamma)\mathbb{E}\Big[r(s_0,a_0)+\ldots r(s_{\tau-1},a_{\tau-1}) + \gamma^\tau \max_\pi \left(E[\sum_{t=0}^{\infty}\gamma^t r(s_{t+\tau},a_{t+\tau})|\pi, s_\tau=s,a_\tau=a]\right) \;\big|\; s_0=s,a_0=a\Big] .
\end{aligned}$$

where the inner max is over all policies which may also use the history of information before time τ. Note that s_τ and a_τ the future evolution at time τ 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 τ, chooses an action that only depends on s_τ. This proves the stationarity claim. Furthermore, by linearity of expectation, the choice of a_τ can made deterministically.

For the second claim, we first show that $Q^\star$ satisfies $Q^\star = \mathcal{T}Q^\star$. We need only consider deterministic policies. We have:

$$\begin{aligned}
Q^\star(s,a) = \max_\pi Q^\pi(s,a) &= \max_\pi \big\{(1-\gamma)r(s.a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^\pi(s')]\big\} \\
&= (1-\gamma)r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[\max_\pi V^\pi(s')] \\
&= (1-\gamma)r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[\max_\pi Q^\pi(s',\pi(s'))] \\
&= (1-\gamma)r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[\max_{\pi,a'} Q^\pi(s',a')] \\
&= (1-\gamma)r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[\max_{a'} Q^\star(s',a')].
\end{aligned}$$

Thus $Q^\star$ satisfies the Bellman optimality equations.

For the converse, suppose $Q = \mathcal{T}Q$ for some Q. For $\pi = \pi_Q$, this implies that $Q = (1-\gamma)r + \gamma P^{\pi_Q} Q$, and so $Q = Q^\pi$, i.e. Q is the action value of the policy π_Q. Now observe for any other policy π':

$$\begin{aligned}
Q^{\pi'} - Q &= (1-\gamma)\left((I-\gamma P^{\pi'})^{-1}r - (I-\gamma P^\pi)^{-1}r\right) \\
&= (I-\gamma P^{\pi'})^{-1}((I-\gamma P^\pi) - (I-\gamma P^{\pi'}))Q^\pi \\
&= \gamma(I-\gamma P^{\pi'})^{-1}(P^{\pi'} - P^\pi)Q^\pi \,.
\end{aligned}$$

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

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

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

1.2 Planning in MDPs

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

1.2.1 Q-Value Iteration

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

$$Q \leftarrow \mathcal{T}Q\,,$$

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

Lemma 1.5. *(contraction) For any two vectors $Q, Q' \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$,*

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

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

$$|V_Q(s) - V_{Q'}(s)| = Q(s,a) - \max_{a' \in \mathcal{A}} Q'(s,a') \leq Q(s,a) - Q'(s,a) \leq \max_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|.$$

Using this,

$$\begin{aligned}
\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty &= \gamma \|PV_Q - PV_{Q'}\|_\infty \\
&= \gamma \|P(V_Q - V_{Q'})\|_\infty \\
&\leq \gamma \|V_Q - V_{Q'}\|_\infty \\
&= \gamma \max_s |V_Q(s) - V_{Q'}(s)| \\
&\leq \gamma \max_s \max_a |Q(s,a) - Q'(s,a)| \\
&= \gamma \|Q - Q'\|_\infty
\end{aligned}$$

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

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

where $\mathbb{1}$ denotes the vector of all ones.

Proof: Fix state s and let $a = \pi_Q(s)$. We have:

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

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

Theorem 1.7. *(Q-value iteration convergence). Set $Q^{(0)} = 0$. For $k = 0, 1, \ldots$, suppose:*

$$Q^{(k+1)} = \mathcal{T} Q^{(k)}$$

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

$$V^{\pi^{(k)}} \geq V^\star - \epsilon \mathbb{1} \, .$$

Proof: Since $\|Q^\star\|_\infty \leq 1$, $Q^{(k)} = \mathcal{T}^k Q^{(0)}$ and $Q^\star = \mathcal{T} Q^\star$, Lemma 1.5 gives

$$\|Q^{(k)} - Q^\star\|_\infty = \|\mathcal{T}^k Q^{(0)} - \mathcal{T}^k Q^\star\|_\infty \leq \gamma^k \|Q^{(0)} - Q^\star\|_\infty = (1-(1-\gamma))^k \|Q^\star\|_\infty \leq \exp(-(1-\gamma)k) \, .$$

The proof is completed with our choice of γ and using Lemma 1.6. ■

1.2.2 Policy Iteration

The policy iteration algorithm starts from an arbitrary policy π_0, and repeat the following iterative procedure: for $k = 0, 1, 2, \ldots$

1. *Policy evaluation.* Compute Q^{π_k}
2. *Policy improvement.* Update the policy:

$$\pi_{k+1} = \pi_{Q^{\pi_k}}$$

In each iteration, we compute the Q-value function of π_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 1.8. *We have that:*

1. $Q^{\pi_{k+1}} \geq \mathcal{T} Q^{\pi_k} \geq Q^{\pi_k}$
2. $\|Q^{\pi_{k+1}} - Q^\star\|_\infty \leq \gamma \|Q^{\pi_k} - Q^\star\|_\infty$

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,

$$
\begin{aligned}
\mathcal{T} Q^{\pi_k}(s,a) &= (1-\gamma) r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_{a'} Q^{\pi_k}(s',a')] \\
&\geq (1-\gamma) r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[Q^{\pi_k}(s', \pi_k(s'))] \\
&= Q^{\pi_k}(s,a).
\end{aligned}
$$

Using this,

$$\begin{aligned}
Q^{\pi_{k+1}}(s,a) &= (1-\gamma)r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[Q^{\pi_{k+1}}(s', \pi_{k+1}(s'))] \\
&\geq (1-\gamma)r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[Q^{\pi_k}(s', \pi_{k+1}(s'))] \\
&= (1-\gamma)r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_{a'} Q^{\pi_k}(s', a')] \\
&= \mathcal{T} Q^{\pi_k}(s,a)
\end{aligned}$$

which proves the first claim.

For the second claim,

$$\|Q^\star - Q^{\pi_{k+1}}\|_\infty \geq \|Q^\star - \mathcal{T} Q^{\pi_k}\|_\infty = \|\mathcal{T} Q^\star - \mathcal{T} Q^{\pi_{k+1}}\|_\infty \leq \gamma \|Q^\star - Q^{\pi_k}\|_\infty$$

where we have used that $Q^\star \geq Q^{\pi_{k+1}} \geq Q^{\pi_k}$ in second step and the contraction property of $\mathcal{T}(\cdot)$ (see Lemma 1.5 in the last step. ■

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

Theorem 1.9. *(policy iteration convergence). Let π_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:*

$$Q^{\pi^{(k)}} \geq Q^\star - \epsilon\,.$$

Chapter 2

Sample Complexity with a Generative Model

Reinforcement Learning and Bandits **Spring 2019**

Sample Complexity with a Generative Model

Instructors: Alekh Agarwal, Sham Kakade *Lecture 2*

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 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 $\widehat{P}$ be our empirical model, defined as follows:

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

where 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 $|\mathcal{S}||\mathcal{A}|N$.

We define $\widehat{M}$ to be the empirical MDP that is identical to the original M, except that it uses $\widehat{P}$ 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 $\widehat{V}^\pi$, $\widehat{Q}^\pi$, $\widehat{Q}^\star$ $\widehat{\pi}^\star$ denote the value function, action value function, and optimal policy in $\widehat{M}$.

2.2 Sample Complexity

2.2.1 A naive approach: accurate model estimation

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

Theorem 2.1. *Let $\epsilon \geq 0$. Suppose we obtain*

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

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:*
$$\max_{s,a} \|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1 \leq (1-\gamma)^2\epsilon/2\,.$$
- *For all policies π,*
$$\|Q^\pi - \widehat{Q}^\pi\|_\infty \leq \epsilon/2$$
- *The estimated $\widehat{Q}^\star$ has error bounded as:*
$$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \epsilon$$

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 $\widehat{Q}^\star$ of $Q^\star$, 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 2.2. *(Crude Value Bounds) Let $\delta \geq 0$. With probability greater than $1-\delta$,*

$$\begin{aligned} \|Q^\star - \widehat{Q}^{\pi^\star}\|_\infty &\leq \Delta_{\delta,N} \\ \|Q^\star - \widehat{Q}^\star\|_\infty &\leq \Delta_{\delta,N}\,, \end{aligned}$$

where:

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

Proof: We have:

$$\begin{aligned} \|Q^\star - \widehat{Q}^{\pi^\star}\|_\infty &= \gamma\|P^{\pi^\star}Q^\star - \widehat{P}^{\pi^\star}\widehat{Q}^{\pi^\star}\|_\infty \\ &\leq \gamma\|P^{\pi^\star}Q^\star - \widehat{P}^{\pi^\star}Q^\star\|_\infty + \gamma\|\widehat{P}^{\pi^\star}Q^\star - \widehat{P}^{\pi^\star}\widehat{Q}^{\pi^\star}\|_\infty \\ &= \gamma\|PV^\star - \widehat{P}V^\star\|_\infty + \gamma\|\widehat{P}^{\pi^\star}(Q^\star - \widehat{Q}^{\pi^\star})\|_\infty \\ &\leq \gamma\|(P-\widehat{P})V^\star\|_\infty + \gamma\|Q^\star - \widehat{Q}^{\pi^\star}\|_\infty\,, \end{aligned}$$

and so we have shown that:

$$\|Q^\star - \widehat{Q}^{\pi^\star}\|_\infty \leq \frac{\gamma}{1-\gamma}\|(P-\widehat{P})V^\star\|_\infty$$

By applying Hoeffding's inequality and the union bound,

$$\|(P-\widehat{P})V^\star\|_\infty = \max_{s,a}|\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^\star(s')] - \mathbb{E}_{s'\sim \widehat{P}(\cdot|s,a)}[V^\star(s')]| \leq \sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}$$

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 $\widehat{Q}^\star$ to be optimal:

Theorem 2.3. *(Azar et al. [2013]) For $\delta \geq 0$ and with probability greater than $1-\delta$,*

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

where c is an absolute constant.

Corollary 2.4. *Let $0 \leq \epsilon \leq \frac{1}{1-\gamma}$. Suppose we obtain*

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

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

$$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \epsilon$$

2.2.3 Lower Bounds

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

Theorem 2.5. *(Azar et al. [2013]) There exists ϵ_0, δ_0, c and a set of MDPs $\mathcal{M}$ such that for $\epsilon \in (0, \epsilon_0)$ and $\delta \in (0, \delta_0)$ if algorithm $\mathcal{A}$ is (ϵ, δ)-good on all $M \in \mathcal{M}$, then $\mathcal{A}$ must use a number of samples that is lower bounded as follows*

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

2.2.4 What about the Value of the Policy $\widehat{\pi}^\star$?

Ultimately, we are interested in the value $V^{\widehat{\pi}^\star}$ when we execute $\widehat{\pi}^\star$, not just an estimate $\widehat{Q}^\star$ of $Q^\star$. The following is an immediate corollary by Lemma 1.6.

Corollary 2.6. *For $\delta \geq 0$ and with probability greater than $1-\delta$,*

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

where c is an absolute constant.

This bound is not sharp. Azar et al. [2013] shows that for sufficiently small ϵ — for $\epsilon \leq c'(1-\gamma)/|\mathcal{S}|$ (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 ϵ 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 2.7. *(Component-wise Bounds) We have that:*

$$\begin{aligned} Q^\star - \widehat{Q}^\star &\leq \gamma (I - \gamma \widehat{P}^{\pi^\star})^{-1} (P - \widehat{P}) V^\star \\ Q^\star - \widehat{Q}^\star &\geq \gamma (I - \gamma \widehat{P}^{\widehat{\pi}^\star})^{-1} (P - \widehat{P}) V^\star \end{aligned}$$

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

$$\begin{aligned} Q^\star - \widehat{Q}^\star &= Q^{\pi^\star} - \widehat{Q}^{\widehat{\pi}^\star} \\ &\leq Q^{\pi^\star} - \widehat{Q}^{\pi^\star} \\ &= (1-\gamma)\left((I - \gamma P^{\pi^\star})^{-1} r - (I - \gamma \widehat{P}^{\pi^\star})^{-1} r\right) \\ &= (I - \gamma \widehat{P}^{\pi^\star})^{-1}((I - \gamma \widehat{P}^{\pi^\star}) - (I - \gamma P^{\pi^\star}))Q^\star \\ &= \gamma (I - \gamma \widehat{P}^{\pi^\star})^{-1}(P^{\pi^\star} - \widehat{P}^{\pi^\star})Q^\star \\ &= \gamma (I - \gamma \widehat{P}^{\pi^\star})^{-1}(P - \widehat{P})V^\star, \end{aligned}$$

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 $\mathcal{D}$ as:

$$\mathrm{Var}_{\mathcal{D}}(f) := E_{x\sim\mathcal{D}}[f(x)^2] - (E_{x\sim\mathcal{D}}[f(x)])^2$$

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

$$\mathrm{Var}_P(V)(s,a) := \mathrm{Var}_{P(\cdot|s,a)}(V)$$

Equivalently,

$$\mathrm{Var}_P(V) = P(V)^2 - (PV)^2 \,.$$

Lemma 2.8. *Let $\delta \geq 0$. With probability greater than $1-\delta$,*

$$|(P-\widehat{P})V^\star| \leq \sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}\sqrt{\mathrm{Var}_P(V^\star)} + \frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{3N}\mathbb{1}\,.$$

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 $\|(I-\gamma\widehat{P}^{\pi^\star})^{-1}\sqrt{\mathrm{Var}_P(V^\star)}\|_\infty$ and $\|(I-\gamma\widehat{P}^{\widehat{\pi}^\star})^{-1}\sqrt{\mathrm{Var}_P(V^\star)}\|_\infty$.

It is helpful to define Σ_M^π as the variance of the discounted reward, i.e.

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

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

The following lemma shows that Σ_M^π satisfies a Bellman consistency condition.

Lemma 2.9. *(Bellman consistency of Σ) For any MDP M,*

$$\Sigma_M^\pi = \gamma^2\mathrm{Var}_P(V_M^\pi) + \gamma^2 P^\pi \Sigma_M^\pi \tag{2.1}$$

where P is the transition model in MDP M.

The proof is left as an exercise to the reader.

Lemma 2.10. *For any policy π and MDP M,*

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

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 ν (where ν is vector of the same dimension of v), Jensen's inequality implies that $\nu\cdot\sqrt{v} \leq \sqrt{\nu\cdot v}$. This implies:

$$\begin{aligned}
\|(I-\gamma P^\pi)^{-1}\sqrt{v}\|_\infty &= \frac{1}{1-\gamma}\|(1-\gamma)(I-\gamma P^\pi)^{-1}\sqrt{v}\|_\infty \\
&\leq \sqrt{\left\|\frac{1}{1-\gamma}(I-\gamma P^\pi)^{-1}v\right\|_\infty} \\
&\leq \sqrt{\left\|\frac{2}{1-\gamma}(I-\gamma^2 P^\pi)^{-1}v\right\|_\infty}\,.
\end{aligned}$$

where we have used that $(I-\gamma P^\pi)^{-1}v \le 2(I-\gamma^2 P^\pi)^{-1}v$ for $v \ge 0$ (which we will prove shortly). The proof is completed as follows: by Equation 2.1, $\Sigma_M^\pi = \gamma^2(I-\gamma^2 P^\pi)^{-1}\mathrm{Var}_P(V_M^\pi)$, so taking $v = \mathrm{Var}_P(V_M^\pi) \ge 0$ and using that $\|\Sigma_M^\pi\|_\infty \le \gamma^2$ completes the proof.

Finally, to see for $v \ge 0$ that $(I-\gamma P^\pi)^{-1}v \le 2(I-\gamma^2 P^\pi)^{-1}v$, observe:

$$\begin{aligned}
2(I-\gamma^2 P^\pi)^{-1}v - (I-\gamma P^\pi)^{-1}v &\ge (1+\gamma)(I-\gamma^2 P^\pi)^{-1}v - (I-\gamma P^\pi)^{-1}v \\
&= (I-\gamma^2 P^\pi)^{-1}\left((1+\gamma)(I-\gamma P^\pi) - (I-\gamma^2 P^\pi)\right)(I-\gamma P^\pi)^{-1}v \\
&= (I-\gamma^2 P^\pi)^{-1}(I-\gamma P^\pi)(I-\gamma P^\pi)^{-1}v \\
&= (I-\gamma^2 P^\pi)^{-1}v \ge 0
\end{aligned}$$

using that $v \ge 0$ and $(I-\gamma^2 P^\pi)^{-1}$ has positive entries. This proves the claim. ∎

Lemma 2.11. *Let $\delta \ge 0$. With probability greater than $1-\delta$, we have:*

$$\mathrm{Var}_P(V^\star) \le 2\mathrm{Var}_{\widehat{P}}(\widehat{V}^{\pi^\star}) + \Delta'_{\delta,N}\mathbb{1}$$
$$\mathrm{Var}_P(V^\star) \le 2\mathrm{Var}_{\widehat{P}}(\widehat{V}^{\star}) + \Delta'_{\delta,N}\mathbb{1}$$

where

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

Proof: By definition,

$$\begin{aligned}
\mathrm{Var}_P(V^\star) &= \mathrm{Var}_P(V^\star) - \mathrm{Var}_{\widehat{P}}(V^\star) + \mathrm{Var}_{\widehat{P}}(V^\star) \\
&= P(V^\star)^2 - (PV^\star)^2 - \widehat{P}(V^\star)^2 + (\widehat{P}V^\star)^2 + \mathrm{Var}_{\widehat{P}}(V^\star) \\
&= (P-\widehat{P})(V^\star)^2 - \left((PV^\star)^2 - (\widehat{P}V^\star)^2\right) + \mathrm{Var}_{\widehat{P}}(V^\star)
\end{aligned}$$

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

$$\|(P-\widehat{P})(V^\star)^2\|_\infty \le \sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}\,.$$

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

$$\|(PV^\star)^2 - (\widehat{P}V^\star)^2\|_\infty \le \|PV^\star + \widehat{P}V^\star\|_\infty \|PV^\star - \widehat{P}V^\star\|_\infty \le 2\|(P-\widehat{P})V^\star\|_\infty \le 2\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}\,.$$

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

$$\begin{aligned}
\mathrm{Var}_{\widehat{P}}(V^\star) &= \mathrm{Var}_{\widehat{P}}(V^\star - \widehat{V}^{\pi^\star} + \widehat{V}^{\pi^\star}) \\
&\le 2\mathrm{Var}_{\widehat{P}}(V^\star - \widehat{V}^{\pi^\star}) + 2\mathrm{Var}_{\widehat{P}}(\widehat{V}^{\pi^\star}) \\
&\le 2\|V^\star - \widehat{V}^{\pi^\star}\|_\infty^2 + 2\mathrm{Var}_{\widehat{P}}(\widehat{V}^{\pi^\star}) \\
&= 2\Delta_{\delta,N}^2 + 2\mathrm{Var}_{\widehat{P}}(\widehat{V}^{\pi^\star})\,.
\end{aligned}$$

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

Using Lemma 2.8 and 2.11, we have the following corollary.

Corollary 2.12. *Let $\delta \geq 0$. With probability greater than $1-\delta$, we have:*

$$|(P-\widehat{P})V^\star| \leq c\sqrt{\frac{\mathrm{Var}_{\widehat{P}}(\widehat{V}^{\pi^\star})\log(c|\mathcal{S}||\mathcal{A}|/\delta)}{N}} + \Delta'_{\delta,N}\mathbb{1}$$
$$|(P-\widehat{P})V^\star| \leq c\sqrt{\frac{\mathrm{Var}_{\widehat{P}}(\widehat{V}^{\star})\log(c|\mathcal{S}||\mathcal{A}|/\delta)}{N}} + \Delta''_{\delta,N}\mathbb{1}\,,$$

where

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

and where c is an absolute constant.

2.3.1 Completing the proof

Proof:(of Theorem 2.3) The proof consists of bounding the terms in Lemma 2.7. We have:

$$\begin{aligned}
& \gamma\|(I-\gamma\widehat{P}^{\pi^\star})^{-1}(P-\widehat{P})V^\star\|_\infty \\
\leq\ & c\gamma\sqrt{\frac{\log(c|\mathcal{S}||\mathcal{A}|/\delta)}{N}}\|(I-\gamma\widehat{P}^{\pi^\star})^{-1}\sqrt{\mathrm{Var}_{\widehat{P}}(\widehat{V}^{\pi^\star})}\|_\infty + \frac{c\gamma}{1-\gamma}\left(\frac{\log(c|\mathcal{S}||\mathcal{A}|/\delta)}{N}\right)^{3/4} \\
& + \frac{c\gamma}{(1-\gamma)^2}\frac{\log(c|\mathcal{S}||\mathcal{A}|/\delta)}{N} \\
\leq\ & \gamma\sqrt{\frac{2}{1-\gamma}}\sqrt{\frac{\log(c|\mathcal{S}||\mathcal{A}|/\delta)}{N}} + \frac{c\gamma}{1-\gamma}\left(\frac{\log(c|\mathcal{S}||\mathcal{A}|/\delta)}{N}\right)^{3/4} + \frac{c\gamma}{(1-\gamma)^2}\frac{\log(c|\mathcal{S}||\mathcal{A}|/\delta)}{N} \\
\leq\ & 3\gamma\sqrt{\frac{1}{1-\gamma}}c\sqrt{\frac{\log(c|\mathcal{S}||\mathcal{A}|/\delta)}{N}} + 2\frac{c}{(1-\gamma)^2}\frac{\log(c|\mathcal{S}||\mathcal{A}|/\delta)}{N}\,,
\end{aligned}$$

where the first step uses Corollary 2.12; the second uses Lemma 2.10; 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. ■

Chapter 3

Strategic Exploration in RL

Reinforcement Learning and Bandits **Spring 2019**

Strategic Exploration in Reinforcement Learning

Instructors: Alekh Agarwal, Sham Kakade *Lecture 3*

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 $s, a, s' \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and $R(s,a) = 0$.
2: Observe initial state s_0 and let π_0 be an arbitrary initial policy.
3: **for all** rounds $t = 0, 1, 2, \ldots$ **do**
4:     **if** a state *has become known*, i.e. $n(s,a) \geq m$ for all $a \in \mathcal{A}$ **then**
5:         update $K = K \cup \{s\}$.
6:         Let $\widehat{M}$ have $\widehat{P}(s'|s,a) = n(s,a,s')/n(s,a)$ and $\widehat{r}(s,a) = R(s,a)/n(s,a)$.
7:         Let $\widehat{M}_K$ be the induced MDP (see Definition 3.1) and $\pi_t = \pi^\star(\widehat{M}_K)$ be the optimal policy in $\widehat{M}_K$.
8:     **else**
9:         If $t \geq 1$, $\pi_t = \pi_{t-1}$.
10:     **end if**
11:     If $s_t \in K$, choose $a_t = \pi_t(s_t)$, else $a_t = \arg\min_{a \in \mathcal{A}} n(s,a)$.
12:     Receive reward r_t and observe next state s_{t+1}.
13:     **if** $s_t \notin K$ **then**
14:         Update $n(s_t, a_t) += 1$, $R(s_t, a_t) += r_t$ and $n(s_t, a_t, s_{t+1}) += 1$.
15:     **end if**
16: **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 7. 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 7 in Algorithm 1 to computing a non-stationary H-step optimal policy instead, where $H = O\left(\frac{\log \frac{1}{\epsilon}}{1-\gamma}\right)$ is the *effective horizon*. 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 ϵ 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 3.1 (Induced MDP)**.** Let M be an MDP parametrized by $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with $K \in \mathcal{S}$ being a subset of states. Based on this set, we define the *induced MDP* M_K parametrized by $(\mathcal{S}, \mathcal{A}, P_{M_K}, r_{M_K}, \gamma)$ 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

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

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 $r_t \in [0,1]$ 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-γ algorithm, and provide a bound on the number of episodes before which it finds an ϵ-optimal policy. For the analysis, it will be useful to invoke the notation

$$H = \frac{\log(2/\epsilon(1-\gamma))}{1-\gamma}.$$

We will prove the following theorem.

Theorem 3.2. *Let s_t be the state visited by the* R-MAX-γ *algorithm at round t and let the parameter m for deciding known states be set as $m = \mathcal{O}\left(\frac{SH^2}{\epsilon^2} \log \frac{S^2 A}{\delta}\right)$. For any $0 \le \epsilon, \delta < 1$, with probability at least $1 - \delta$, $V_M^{\pi_t}(s_t) \ge V_M^\star(s_t) - \epsilon$, for all but $\mathcal{O}\left(\frac{H^3 S^2 A}{\epsilon^3} \log \frac{S^2 A}{\delta}\right)$ rounds 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 $\mathcal{O}\left(\frac{H^3 S^2 A}{\epsilon^3} \log \frac{S^2 A}{\delta}\right)$ 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 ϵ 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 3.3 (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*

$$\sum_{s' \in \mathcal{S}} |P_M(s'|s,a) - P_{M'}(s'|s,a)| \leq \epsilon_1, \quad \forall s \in \mathcal{S} \text{ and } a \in \mathcal{A}, |r_M(s,a) - r_{M'}(s,a)| \leq \epsilon_2 \quad \forall s \in \mathcal{S} \text{ and } a \in \mathcal{A}.$$

Then for every stationary policy π, the two MDPs satisfy $\|V^\pi_M - V^\pi_{M'}\|_\infty \leq \frac{\gamma}{1-\gamma}\epsilon_1 + \epsilon_2$.

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

$$\begin{aligned}
|V^\pi_M(s) - V^\pi_{M'}(s)| &\leq (1-\gamma)\epsilon_2 + \gamma \left| \sum_{s' \in \mathcal{S}} (P_M(s'|s,\pi(s)) V^\pi_M(s') - P_{M'}(s'|s,\pi(s)) V^\pi_{M'}(s')) \right| \\
&\leq (1-\gamma)\epsilon_2 + \gamma \left| \sum_{s' \in \mathcal{S}} P_M(s'|s,a)(V^\pi_M(s') - V^\pi_{M'}(s')) \right| + \gamma \left| \sum_{s' \in \mathcal{S}} V^\pi_{M'}(s')(P_M(s'|s,a) - P_{M'}(s'|s,a)) \right| \\
&\leq (1-\gamma)\epsilon_2 + \gamma \|V^\pi_M - V^\pi_{M'}\|_\infty + \gamma\epsilon_1.
\end{aligned}$$

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 π_i learned using the induced MDP at an epsisode i is already good, or it has a high chance of taking us to an unknown state. We will use the notation

$$\mathbb{P}^\pi_M[\text{escape from } K | s_0 = s] := \mathbf{1}(s \notin K) + \sum_{t=1}^{\infty} \gamma^t \mathbb{P}^\pi_M(s_t \notin K, s_0, \ldots, s_{t-1} \in K).$$

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

Lemma 3.4 (Induced inequalities)**.** *Let M be an MDP with K being the set of known states. Let M_K be the induced MDP (Definition 3.1) with respect to K and M. For any stationary policy π and state $s \in \mathcal{S}$ we have*

$$V^\pi_{M_K}(s) \geq V^\pi_M(s) \quad \textit{and} \quad V^\pi_M(s) \geq V^\pi_{M_K}(s) - \mathbb{P}^\pi_M[\textit{escape from } K | s_0 = s].$$

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 π 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

$$\begin{aligned}
|V_M^\pi(s) - V_{M_K}^\pi(s)| &\le \mathbf{1}(s \notin K) + \mathbf{1}(s \in K)\gamma \left|\sum_{s' \in \mathcal{S}} P_M(s'|s,\pi(s))V_M^\pi(s') - P_{M_K}(s'|s,\pi(s))V_{M_K}^\pi(s')\right| \\
&= \mathbf{1}(s \notin K) + \mathbf{1}(s \in K)\gamma \left|\sum_{s' \in \mathcal{S}} P_M(s'|s,\pi(s))(V_M^\pi(s') - V_{M_K}^\pi(s'))\right| \\
&\le \mathbf{1}(s \notin K) + \mathbf{1}(s \in K)\gamma P_M(s' \notin K|s,\pi(s)) + \mathbf{1}(s \in K)\gamma \left|\sum_{s' \in K} P_M(s'|s,\pi(s))(V_M^\pi(s') - V_{M_K}^\pi(s'))\right|.
\end{aligned}$$

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 π_i computed in each episode of Algorithm 1 is near optimal, with the error being the probability of leaving the known state set.

Corollary 3.5 (Implicit Explore-Exploit)**.**

$$V_M^{\pi^\star(M_K)}(s) \ge V_M^\star(s) - \mathbb{P}_M^{\pi^\star(M_K)}[\textit{escape from } K | s_0 = s]$$

Proof: By the lemma, we have

$$\begin{aligned}
V_M^{\pi^\star(M_K)}(s) &\ge V_{M_K}^{\pi^\star(M_K)}(s) - \mathbb{P}_M^{\pi^\star(M_K)}[\text{escape from } K | s_0 = s] \\
&\ge V_{M_K}^{\pi^\star(M)}(s) - \mathbb{P}_M^{\pi^\star(M_K)}[\text{escape from } K | s_0 = s] \\
&\ge V_{M}^{\pi^\star(M)}(s) - \mathbb{P}_M^{\pi^\star(M_K)}[\text{escape from } K | s_0 = s].
\end{aligned}$$

Here the first inequality follows from Lemma 3.4 applied with $\pi = \pi^\star(M_K)$, second inequality uses that $\pi^\star(M_K)$ is the optimal policy in M_K and hence obtains a higher reward than $\pi^\star(M)$ and the third inequality follows from the optimism of M_K shown in Lemma 3.4. ■

We conclude with a final lemma which relates the probability of escape from K over an infinite trajectory to that of encountering an unknown state over H steps.

Lemma 3.6. *With probability at least $1 - \delta$, the number of rounds where $V_M^{\pi_t}(s_t) \le V_M^\star(s_t) - \epsilon$ is at most $\mathcal{O}\left(\frac{mHSA}{\epsilon} \ln \frac{1}{\delta}\right)$.*

Proof: In this lemma, we with sufficiently many visits to states with a large escape probability, all the states become known. Furthermore, with high probability, the number of rounds following such a visit where our policy's value function is significantly suboptimal is at most H.

The proof of the lemma has two parts. First we show that if at round t, the probability of escape from K starting from s_t is large, then we also have a large probability of escape in the next H steps. Using Bernoulli concentration, we then further show that the algorithm will indeed encounter an unknown state in the next H steps with a large probability. For the first part, let $s = s_t$ be the state encountered at some round t and let $\pi = \pi_t$ be the current policy. Suppose further that we know $\mathbb{P}_M^\pi[\text{escape from } K \mid s_0 = s] \ge \epsilon$. Let us define

$$p_H = \mathbf{1}(S \notin K) + \sum_{t=1}^{H} \mathbb{P}_M^\pi(s_t \notin K \mid s_0, \ldots, s_{t-1} \in K).$$

Note that there is no discounting in the definition of p_H. We have

$$\begin{aligned}
\epsilon &\le \mathbf{1}(s \notin K) + \sum_{t=1}^{\infty} \gamma^t \mathbb{P}_M^\pi(s_t \notin K \mid s_0, \ldots, s_{t-1} \in K) \\
&\le p_H + \sum_{t=H+1}^{\infty} \gamma^t \mathbb{P}_M^\pi(s_t \notin K \mid s_0, \ldots, s_{t-1} \in K) \\
&\le p_h + \frac{\gamma^{H+1}}{1-\gamma}.
\end{aligned}$$

Thus we see that $p_H \ge \epsilon - \gamma^{H+1}/(1-\gamma)$. It suffices to ensure that the tail term is at most $\epsilon/2$ to guarantee that $p_H \ge \epsilon/2$, which means that we want $H \ge \log \frac{2}{\epsilon(1-\gamma)} / \log \frac{1}{\gamma}$. Noting that $\log(1/\gamma) \ge 1-\gamma$ for $\gamma \in (0,1]$ completes a lower bound of $\epsilon/2$ on p_H for the stated value of H.

It remains to bound the number of actions before we have sufficiently many visits to unknown states. For this, let $t_1, t_2, \ldots$ be the rounds such that $|t_i) - t_{i+1}| > H$ and if π_i is the policy used at round t_i and K_i is the set of known states at t_i, then $\mathbb{P}^{\pi_i}(\text{escape from } K \mid s_0 = s_{t_i}) \ge \epsilon$. Let us define a random variable

$$X_i = \mathbf{1}(\exists s \{s_{t_i}, s_{t_i+1}, \ldots, s_{t_i+H}\} \;:\; s \notin K_i).$$

By the definition of the rounds t_i, we know that $\mathbb{E}[X_i | s_{t_i}] \ge \epsilon$. Let us define a σ-field $\mathcal{F}_i$ to consist of all the randomness in the MDP and the agent prior to t_i (this includes the randomness in the state transitions at round $t_{i+1}-1$, so that the state $s_{t_{i+1}}$, the policy π_{i+1} and the known states K_{i+1} are known when conditioning on $\mathcal{F}_i$). It is easily checked that X_i is measurable with respect to $\mathcal{F}_i$, since the entire trajectory till round $t_{i+1}-1$ is known under this conditioning, which in particular implies that we know the subtrajectory from rounds t_i to $t_i + H$ and hence can check the indicator in X_i. Also, we observe that $\mathbb{E}[X_i \,|\, \mathcal{F}_{i-1}] = \mathbb{P}^{\pi_i}(\text{escape from } K_i \text{ in } H \text{ steps} \,|\, s_{t_i}) \ge \epsilon/2$ using the earlier lower bound on p_H. Furthermore, since the X_i are binary valued, it is clear that

$$\mathbb{E}[(X_i - \mathbb{E}[X_i \mid \mathcal{F}_{i-1}])^2 \mid \mathcal{F}_{i-1}] \le \mathbb{E}[X_i^2 \mid \mathcal{F}_{i-1}] \le \mathbb{E}[X_i \mid \mathcal{F}_{i-1}].$$

Now applying the Freedman's inequality for martingales (Lemma A.3 in the appendix) to the sequence $Y_i = X_i - \mathbb{E}[X_i \mid \mathcal{F}_{i-1}]$, we see that with probability at least $1-\delta$, for a fixed n we have

$$\begin{aligned}
\sum_{i=1}^{n} (\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] - X_i) &\le 2\sqrt{\ln\frac{1}{\delta} \sum_{i=1}^{n} \mathbb{E}[X_i \mid \mathcal{F}_{i-1}] + \ln\frac{1}{\delta}} \\
&\le \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}[X_i \mid \mathcal{F}_{i-1}] + 3\ln\frac{1}{\delta}.
\end{aligned}$$

This implies that

$$\sum_{i=1}^{n} X_i \ge \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}[X_i \mid \mathcal{F}_{i-1}] - 3\ln\frac{1}{\delta} \ge \frac{n\epsilon}{4} - 3\ln\frac{1}{\delta}.$$

Since we desire $\sum_{i=1}^{n} X_i \ge mSA$, it suffices to pick $n \ge 4(mSA + 3\ln(1/\delta))/\epsilon$.

Note that for the rounds $t \in [t_{i+H+1}, t_{i+1} - 1]$, we already know that the probability of escape is small, meaning that the value function of our policy is near optimal on those rounds. This concludes the proof. ■

Proof of Theorem 3.2 .

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 3.3, 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

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

Combining with Lemma 3.4, we see that for any starting state we have

$$\begin{aligned} V_M^{\pi_t}(s) &\geq V_{M_K}^{\pi^\star}(s) - \epsilon - \mathbb{P}_M^{\pi_t}[\text{escape from } K | s_0 = s] \\ &\geq V_M^{\pi^\star}(s) - \epsilon - \mathbb{P}_M^{\pi_t}[\text{escape from } K | s_0 = s], \end{aligned}$$

where the first inequality is combining Lemma 3.4 with with the earlier bound, and the second inequality uses the optimism of the induced MDP M_K.

Thus, either the policy π_t is at most 2ϵ suboptimal, or it visits an unknown state with probability at least ϵ. Intuitively, this means that we visit an unknown state at least every $1/\epsilon$ steps, if $\widehat{\pi}_t$ is not already near optimal. Since the probabilities are discounted,

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. This intuition is made precise in Lemma 3.6.

In order to obtain the theorem statement, we set $m = \mathcal{O}\left(\frac{SH^2}{\epsilon^2} \log \frac{S^2 A}{\delta}\right)$. This number is based on satisfying the condition of Lemma 3.3 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.

Chapter 4

Policy Gradient Methods

Reinforcement Learning and Bandits **Spring 2019**

Policy Gradient Methods

Instructors: Alekh Agarwal, Sham Kakade *Lecture 4*

We now work with an initial state distribution d_0 where $s_0 \sim d_0$. Slightly overloading notation, let us define:

$$V_M^\pi = V_M^\pi(d_0) := \mathbb{E}_{s_0 \sim d_0}[V_M^\pi(s_0)]\,,$$

where we drop the d_0 dependence when clear from context. Consider a class of parametric policies $\{\pi_\theta | \theta \in \Theta \subset \mathbb{R}^d\}$. The optimization problem of interest is:

$$\max_{\theta \in \Theta} V_M^{\pi_\theta}\,.$$

The most common approach is to use an interior point, continuous optimization algorithm, such as gradient ascent. The hope of a "direct" approach is that it may more easily deal with large state and action spaces, and, possibly, more readily handle settings where the model is not known. One immediate issue is that if the policy class $\{\pi_\theta\}$ consists of deterministic policies then π_θ will, in general, not be differentiable. This motivates us to consider policy classes that are stochastic, which permit differentiability. For example, $\pi_\theta(a|s)$ may be parameterized by some neural network.

It is instructive to explicitly consider a "tabular" policy representation, given by the *Gibbs policy*:

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_a \exp(\theta_{s,a})}\,, \tag{4.1}$$

where the parameter space is $\Theta = \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Note that (the closure of) the set of Gibbs policies contains all stationary and deterministic policies.

Advantages and the state-action visitation distribution: Let us first introduce the concept of an advantage.

Definition 4.1. The advantage $A_M^\pi(s,a)$ of a policy π for an MDP M is defined as

$$A_M^\pi(s,a) := Q_M^\pi(s,a) - V_M^\pi(s)\,.$$

Note that:

$$A_M^*(s,a) := A_M^{\pi^*}(s,a) \le 0$$

for all state-action pairs.

We now define the discounted state visitation distribution. Let τ denote a trajectory, whose unconditional distribution $\Pr_M^\pi(\tau)$ under π in MDP M is

$$\Pr{}_M^\pi(\tau) = d_0(s_0)\pi(a_0|s_0)P(s_1|s_0,a_0)\pi(a_1|s_1)\cdots\,. \tag{4.2}$$

Our notation can be simplified through defining the discounted state visitation distribution $d_{s_0}^\pi$ as:

$$d_{s_0}^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr{}^\pi(s_t = s|s_0)$$

where $\Pr^\pi(s_t = s|s_0)$ is the state-action visitation probability, where we use π in M starting at state s_0. We also write:

$$d^\pi(s) = d_{d_0}^\pi(s) = \mathbb{E}_{s_0 \sim d_0}\left[d_{s_0}^\pi(s)\right]$$

where we drop the d_0 subscript when clear from context.

Lemma 4.2. *(The performance difference lemma) For all policies π, π' and states s,*

$$\begin{aligned} V^{\pi}(s) - V^{\pi'}(s) &= \mathbb{E}_{\tau \sim \Pr^{\pi}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi'}(s_t, a_t)\right] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_s^{\pi}} \mathbb{E}_{a' \sim \pi(\cdot|s')}\left[A^{\pi'}(s', a')\right] \end{aligned}$$

Proof: Using a telescoping argument, we have:

$$\begin{aligned} V^{\pi}(s) - V^{\pi'}(s) &= (1-\gamma)\mathbb{E}_{\tau \sim \Pr^{\pi}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] - V^{\pi'}(s) \\ &= \mathbb{E}_{\tau \sim \Pr^{\pi}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty} \gamma^t \left((1-\gamma) r(s_t, a_t) + V^{\pi'}(s_t) - V^{\pi'}(s_t)\right)\right] - V^{\pi'}(s) \\ &= \mathbb{E}_{\tau \sim \Pr^{\pi}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty} \gamma^t \left((1-\gamma) r(s_t, a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t)\right)\right] \\ &\overset{(a)}{=} \mathbb{E}_{\tau \sim \Pr^{\pi}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty} \gamma^t \left((1-\gamma) r(s_t, a_t) + \gamma \mathbb{E}[V^{\pi'}(s_{t+1})|s_t, a_t] - V^{\pi'}(s_t)\right)\right] \\ &= \mathbb{E}_{\tau \sim \Pr^{\pi}(\tau|s_0=s)}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi'}(s_t, a_t)\right] \end{aligned}$$

where (a) uses the tower property of conditional expectations. ■

4.1 The Policy Gradient Method

Analytical form: It is convenient to define the discounted total reward of a trajectory as:

$$R(\tau) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

where s_t, a_t are the state-action pairs in τ. Observe that:

$$V^{\pi_\theta} = \mathbb{E}_{\tau \sim \Pr^{\pi_\theta}}\left[R(\tau)\right].$$

Theorem 4.3. *(Policy gradients) The following are expressions for $\nabla_\theta V^{\pi_\theta}$:*

- *REINFORCE [Williams, 1992]:*

$$\nabla V^{\pi_\theta} = \mathbb{E}_{\tau \sim \Pr^{\pi_\theta}}\left[R(\tau) \sum_{t=0}^{\infty} \nabla \log \pi_\theta(a_t|s_t)\right]$$

- *Action value expression:*

$$\begin{aligned} \nabla V^{\pi_\theta} &= \mathbb{E}_{\tau \sim \Pr^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t, a_t) \nabla \log \pi_\theta(a_t|s_t)\right] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}\left[Q^{\pi_\theta}(s, a) \nabla \log \pi_\theta(a|s)\right] \end{aligned}$$

- *Advantage expression:*

$$\nabla V^{\pi_\theta} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\Big[A^{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)\Big]$$

The alternative expressions are more helpful to use when we turn to Monte Carlo estimation.

Proof: We have:

$$\begin{aligned}
\nabla V^{\pi_\theta} &= \nabla\sum_\tau R(\tau)\mathrm{Pr}^{\pi_\theta}(\tau)\\
&= \sum_\tau R(\tau)\nabla\mathrm{Pr}^{\pi_\theta}(\tau)\\
&= \nabla\sum_\tau R(\tau)\mathrm{Pr}^{\pi_\theta}(\tau)\nabla\log\mathrm{Pr}^{\pi_\theta}(\tau)\\
&= \sum_\tau R(\tau)\mathrm{Pr}^{\pi_\theta}(\tau)\nabla\log\left(d_0(s_0)\pi_\theta(a_0|s_0)P(s_1|s_0,a_0)\pi_\theta(a_1|s_1)\cdots\right)\\
&= \sum_\tau R(\tau)\mathrm{Pr}^{\pi_\theta}(\tau)\left(\sum_{t=0}^{\infty}\nabla\log\pi_\theta(a_t|s_t)\right)
\end{aligned}$$

which completes the proof of the first claim.

For the second claim, for any state s_0

$$\begin{aligned}
&\nabla V^{\pi_\theta}(s_0)\\
=\ &\nabla\sum_{a_0}\pi_\theta(a_0|s_0)Q^{\pi_\theta}(s_0,a_0)\\
=\ &\sum_{a_0}\Big(\nabla\pi_\theta(a_0|s_0)\Big)Q^{\pi_\theta}(s_0,a_0)+\sum_{a_0}\pi_\theta(a_0|s_0)\nabla Q^{\pi_\theta}(s_0,a_0)\\
=\ &\sum_{a_0}\pi_\theta(a_0|s_0)\Big(\nabla\log\pi_\theta(a_0|s_0)\Big)Q^{\pi_\theta}(s_0,a_0)\\
&+\sum_{a_0}\pi_\theta(a_0|s_0)\nabla\Big((1-\gamma)r(s_0,a_0)+\gamma\sum_{s_1}P(s_1|s_0,a_0)V^{\pi_\theta}(s_1)\Big)\\
=\ &\sum_{a_0}\pi_\theta(a_0|s_0)\Big(\nabla\log\pi_\theta(a_0|s_0)\Big)Q^{\pi_\theta}(s_0,a_0)+\gamma\sum_{a_0,s_1}\pi_\theta(a_0|s_0)P(s_1|s_0,a_0)\nabla V^{\pi_\theta}(s_1)\\
=\ &\mathbb{E}_{\tau\sim\mathrm{Pr}^{\pi_\theta}(\tau|s_0=s)}\left[Q^{\pi_\theta}(s_0,a_0)\nabla\log\pi_\theta(a_0|s_0)\right]+\gamma\mathbb{E}_{\tau\sim\mathrm{Pr}^{\pi_\theta}(\tau|s_0=s)}\left[\nabla V^{\pi_\theta}(s_1)\right]\\
=\ &\mathbb{E}_{\tau\sim\mathrm{Pr}^{\pi_\theta}(\tau|s_0=s)}\left[Q^{\pi_\theta}(s_0,a_0)\nabla\log\pi_\theta(a_0|s_0)\right]+\gamma\mathbb{E}_{\tau\sim\mathrm{Pr}^{\pi_\theta}(\tau|s_0=s)}\left[Q^{\pi_\theta}(s_1,a_1)\nabla\log\pi_\theta(a_1|s_1)\right]+\ldots.
\end{aligned}$$

where the last step follows from recursion. Taking an expectation over s_0 completes the proof of the second claim.

The proof of the final claim is left as an exercise to the reader. ■

4.1.1 Optimization

Gradient ascent and convergence to stationary points: Let us say a function $f:\mathbb{R}^d\to\mathbb{R}$ is L *-smooth* if

$$\|\nabla f(w)-\nabla f(w')\|\le L\|w-w'\|\,,$$

where the norm $\|\cdot\|$ is the Euclidean norm. In other words, the derivatives of f do not change too quickly.

Gradient ascent, with a stepsize, follows the update rule:

$$\theta_{t+1} = \theta_t + \eta \nabla V^{\pi_\theta} \,.$$

The next lemma is standard in non-convex optimization (e.g. see [Jain and Kar, 2017, Bubeck, 2015]).

Lemma 4.4. *(Convergence to Stationary Points) Assume that for all $\theta \in \Theta$, V^{π_θ} is L-smooth and bounded below by V_*. Suppose we use the constant stepsize $\eta = 1/L$. For all T, we have that*

$$\min_{t\le T} \|\nabla V^{\pi_{\theta_t}}\|^2 \le \frac{2L(V_* - V^{\pi_{\theta_0}})}{T} \,.$$

Monte Carlo estimation and stochastic gradient ascent: One difficulty is that even if we know the MDP M, computing the gradient may be computationally intensive. It turns out that we can obtain unbiased estimates of π with only simulation based access to our model, i.e. assuming we can obtain sampled trajectories $\tau \sim \Pr^{\pi_\theta}$

With respect to a trajectory τ, define:

$$\begin{aligned}
\widehat{Q}_t &:= (1-\gamma) \sum_{t'=t}^{\infty} \gamma^{t'-t} r(s_{t'}, a_{t'}) \\
\widehat{\nabla V^{\pi_\theta}} &:= \sum_{t=0}^{\infty} \gamma^t \widehat{Q}_t \nabla \log \pi_\theta(a_t|s_t)
\end{aligned}$$

We now see that this provides an unbiased estimated of the gradient:

Lemma 4.5. *(Unbiased gradient estimate) We have :*

$$\mathbb{E}_{\tau\sim\Pr^{\pi_\theta}}\left[\widehat{\nabla V^{\pi_\theta}}\right] = \nabla V^{\pi_\theta}$$

Proof: We have:

$$\begin{aligned}
\mathbb{E}[\widehat{\nabla V^{\pi_\theta}}] &= \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \widehat{Q}_t \nabla \log \pi_\theta(a_t|s_t)] \\
&\overset{(a)}{=} \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\widehat{Q}_t|s_t, a_t] \nabla \log \pi_\theta(a_t|s_t)] \\
&\overset{(b)}{=} \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t Q_t(s_t, a_t) \nabla \log \pi_\theta(a_t|s_t)]
\end{aligned}$$

where (a) follows from the tower property of the conditional expectations and (b) follows from the Markov property implying that $\mathbb{E}[\widehat{Q}_t|s_t, a_t] = Q(s_t, a_t)$. ■

The stochastic gradient ascent algorithm is as follows:

1. initialize θ_0.
2. For $t = 0, 1, \ldots$
   (a) Sample $\tau \sim \Pr^{\pi_\theta}$.

(b) Update:

$$\theta_{t+1} = \theta_t + \eta_t \widehat{\nabla V^{\pi_\theta}}$$

where η_t is the stepsize and $\widehat{\nabla V^{\pi_\theta}}$ estimated with τ.

Note here that we are ignoring that τ is an infinte length sequence. It can be truncated appropriately so as to control the bias.

The following is standard result with regards to non-convex optimization [Ghadimi and Lan, 2013, Jain and Kar, 2017]. Again, with reasonably bounded variance, we will obtain a point θ_t with small gradient norm.

Lemma 4.6. *(Stochastic Convergence to Stationary Points) Assume that for all $\theta \in \Theta$, V^{π_θ} is L-smooth and bounded below by V_*. Suppose the variance is bounded as:*

$$\mathbb{E}[\|\widehat{\nabla V^{\pi_\theta}} - \nabla V^{\pi_\theta}\|^2] \leq \sigma^2$$

For $t \leq L(V^{\pi_{\theta_0}} - V_*)/\sigma^2$, suppose we use a constant stepsize of $\eta_t = 1/L$, and then $\eta_t = \sqrt{2/(LT)}$ after. For all T, we have:

$$\min_{t \leq T} \mathbb{E}[\|\nabla V^{\pi_{\theta_t}}\|^2] \leq \frac{2L(V^{\pi_{\theta_0}} - V_*)}{T} + \sqrt{\frac{2\sigma^2}{T}}.$$

Monte Carlo estimation and stochastic gradient ascent: A significant practical issue is that the variance σ^2 is often large in practice. Here, a form of variance reduction is often critical in practice. A common method is as follows.

1. Draw multiple trajectories $\tau_1 \dots \tau_N$, where $\tau_i \sim \Pr^{\pi_\theta}$. Then construct a function $f : \mathcal{S} \to \mathbb{R}$ based on these samples. Often we construct f as an estimate of V^{π_θ}
2. Independently sample another trajectory τ, and define:

$$\begin{aligned}
\widehat{Q}_t &:= (1-\gamma)\sum_{t'=t}^{\infty} \gamma^{t'-t} r(s_{t'}, a_{t'}) \\
\widehat{\nabla V^{\pi_\theta}} &:= \sum_{t=0}^{\infty} \gamma^t \Big(\widehat{Q}_t - f(s_t)\Big) \nabla \log \pi_\theta(a_t|s_t)
\end{aligned}$$

We often refer to $f(s)$ as a baseline at state s. Often this method is coupled with the natural policy gradient algorithm, which we discuss in the next section.

Lemma 4.7. *(Unbiased gradient estimate with Variance Reduction) For any procedure used to construct to the baseline function $f : \mathcal{S} \to \mathbb{R}$, if the samples used to construct f are independent of $\widehat{Q}_t$ then:*

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(\widehat{Q}_t - f(s_t)\Big) \nabla \log \pi_\theta(a_t|s_t)\right] = \nabla V^{\pi_\theta}$$

where the expectation is with respect to both the random trajectory τ and the random function $f(\cdot)$.

Proof: For any function $g(s)$,

$$\mathbb{E}\left[\nabla \log \pi(a|s) g(s)\right] = \sum_a \nabla \pi(a|s) g(s) = g(s) \sum_a \nabla \pi(a|s) = g(s) \nabla \sum_a \pi(a|s) = g(s) \nabla 1 = 0$$

Using that $f(\cdot)$ is independent of τ, we have that for all t

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t f(s_t)\nabla\log\pi_\theta(a_t|s_t)\right]=0$$

The result now follow froms Lemma 4.5. ∎

4.2 Global Convergence of the Gibbs Policy and Entropy Regularization

Let us now return to the Gibbs policy, from Equation 4.1. Even for this case, our optimization problem is non-convex:

Lemma 4.8. *For the Gibbs policy class, there exists an MDP M such that $V_M^{\pi_\theta}$ is not a convex function in θ.*

We leave the proof as an exercise to the reader.

Policy Gradient Expression Observe that:

$$\frac{\partial\log\pi_\theta(a|s)}{\partial\theta_{s',a'}}=\mathbb{1}\Big[s=s'\Big]\Big(\mathbb{1}\Big[a=a'\Big]-\pi_\theta(a'|s)\Big)$$

where $\mathbb{1}[\mathcal{E}]$ is the indicator of $\mathcal{E}$ being true.

Lemma 4.9. *For the Gibbs policy class, we have:*

$$\frac{\partial V^{\pi_\theta}}{\partial\theta_{s,a}}=\frac{1}{1-\gamma}d^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi_\theta}(s,a)$$

Proof: We have:

$$\begin{aligned}
\frac{\partial V^{\pi_\theta}}{\partial\theta_{s,a}} &= \mathbb{E}_{\tau\sim\Pr^{\pi_\theta}}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{1}[s_t=s]\Big(\mathbb{1}[a_t=a]A^{\pi_\theta}(s,a)-\pi_\theta(a|s)A^{\pi_\theta}(s_t,a_t)\Big)\right]\\
&= \mathbb{E}_{\tau\sim\Pr^{\pi_\theta}}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{1}[(s_t,a_t)=(s,a)]A^{\pi_\theta}(s,a)\right]+\pi_\theta(a|s)\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{\tau\sim\Pr^{\pi_\theta}}\left[\mathbb{1}[s_t=s]A^{\pi_\theta}(s_t,a_t)\right]\\
&= \frac{1}{1-\gamma}\mathbb{E}_{(s',a')\sim d^{\pi_\theta}}\left[\mathbb{1}[(s',a')=(s,a)]A^{\pi_\theta}(s,a)\right]+0\\
&= \frac{1}{1-\gamma}d^{\pi_\theta}(s,a)A^{\pi_\theta}(s,a)\,,
\end{aligned}$$

where the second to last step uses that for any policy $\sum_a\pi(a|s)A^\pi(s,a)=0$. ∎

Convergence to the Global Optima. While our optimization problem is not convex, we now see that global convergence to the optimal solution is possible provided we make appropriate algorithmic modifications. The difficulty is that the gradient, with respect to $\theta_{s,a}$, may be small when $\pi_\theta(a|s)$ is small, even if the advantage $A^{\pi_\theta}(s,a)$ is positive and large.

In order to prevent probabilities from becoming arbitrarily small, we consider an entropic regularization penalty. Recall that the cross entropy for distributions p and q is defined as:

$$H(p,q):=\mathbb{E}_{x\sim p}\left[\log\frac{1}{q(x)}\right]$$

Denote the uniform distribution over the action space by Uni, and define the following entropy based regularized objective:

$$\begin{aligned} L_\lambda(\theta) &:= V^{\pi_\theta} - \lambda \sum_s H(\text{Uni}, \pi_\theta(\cdot|s)) \\ &= V^{\pi_\theta} + \frac{\lambda}{|\mathcal{A}|} \sum_{s,a} \log \pi_\theta(a|s) \,, \end{aligned}$$

where λ is a parameter to be set appropriately.

The following theorem shows that with sufficiently accurate optimization (i.e. an appropriately small gradient norm) on the regularized objective, along with an appropriately small regularizer, then the policy gradient algorithm (quickly) converges to a near optimal policy.

Theorem 4.10. *(Global convergence of Gibbs) For the Gibbs policy class, suppose we find a point θ such that:*

$$\|\nabla L_\lambda(\theta)\|_2 \le \epsilon_{optim} \,.$$

Assuming that $\epsilon_{optim} \le \lambda/(2|\mathcal{A}|)$, we have:

$$\begin{aligned} V^{\pi_\theta} &\ge V^* - 4 \left\| \frac{d^{\pi^*}}{d^{\pi_\theta}} \right\|_\infty |\mathcal{A}|\, |\mathcal{S}|\, \lambda \\ &\ge V^* - \frac{4}{1-\gamma} \left\| \frac{d^{\pi^*}}{d_0} \right\|_\infty |\mathcal{A}|\, |\mathcal{S}|\, \lambda \end{aligned}$$

where $\frac{d^{\pi^*}}{d^{\pi_\theta}}$ represents componentwise division.

The above theorem shows the importance of having an appropriate measure $d_0(s)$ in order for a policy gradient method to quickly reach the global optima.

Proof: By the performance difference lemma (Lemma 4.2),

$$\begin{aligned} V^* - V^{\pi_\theta} &= \frac{1}{1-\gamma} \sum_{s,a} d^{\pi^*}(s) \pi^*(a|s) A^{\pi_\theta}(s,a) \\ &\le \frac{1}{1-\gamma} \sum_s d^{\pi^*}(s) \max_{a \in \mathcal{A}} A^{\pi_\theta}(s,a) \\ &\le \frac{1}{1-\gamma} \left\| \frac{d^{\pi^*}}{d^{\pi_\theta}} \right\|_\infty \sum_s d^{\pi_\theta}(s) \max_{a \in \mathcal{A}} A^{\pi_\theta}(s,a) \,. \end{aligned}$$

We will now show that $d^{\pi_\theta}(s) A^{\pi_\theta}(s,a) \le 2(1-\gamma)\lambda$ for every state-action pair, which proves the desired result. Due to that

$$L_\lambda(\theta) = V^{\pi_\theta} + \frac{\lambda}{|\mathcal{A}|} \sum_{s,a} \left(\theta_{s,a} - \log \left(\sum_{a'} \exp(\theta_{s,a}) \right) \right) \,,$$

we have

$$\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d^{\pi_\theta}(s) \pi_\theta(a|s) A^{\pi_\theta}(s,a) + \lambda \left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s) \right) \,.$$

Consider any state-action pair s, a where $A^{\pi_\theta}(s,a) \ge 0$. If there exists no such pair, then π_θ is optimal and the claim is trivially true. The gradient norm assumption implies that:

$$\epsilon_{\text{optim}} \geq \frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi_\theta}(s,a) + \lambda\left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s)\right) \geq \lambda\left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s)\right)$$

Rearranging and using our assumption on ϵ_{optim}, we have

$$\pi_\theta(a|s) \geq \frac{1}{|\mathcal{A}|} - \frac{\epsilon_{\text{optim}}}{\lambda} \geq \frac{1}{2|\mathcal{A}|}\,.$$

With this,

$$\begin{aligned} d^{\pi_\theta}(s)A^{\pi_\theta}(s,a) &= \frac{1-\gamma}{\pi_\theta(a|s)}\left(\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} + \lambda\left(\pi_\theta(a|s) - \frac{1}{|\mathcal{A}|}\right)\right) \\ &\leq (1-\gamma)\left(\frac{1}{\pi_\theta(a|s)}\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} + \lambda\right) \\ &\leq (1-\gamma)\left(2|\mathcal{A}|\epsilon_{\text{optim}} + \lambda\right) \\ &\leq 2\lambda\,. \end{aligned}$$

which completes the proof of our claim after substitution. ■

4.3 Approximation, Optimality, and the Belman Policy Error

What can we say about the quality of the policy we find with gradient ascent? For a restrictive policy class, It is unreasonable to expect to be able to efficiently find a policy whose value is near to that of $\max_{\theta\in\Theta} V^{\pi_\theta}$ due to this being a nonconvex optimization problem. On the other hand, with a rich enough policy class, we may hope that a (near) globally optimal policy will be reached if our optimization is successful, as we saw in the case for Gibbs policy. We make this precise, where we see that the optimality will depend on both a notion of approximation quality and a notion of measure.

To gain intuition, consider the implication of Taylor's theorem on $\theta + \delta$:

$$\pi_{\theta+\delta}(a|s) = \pi_\theta(a|s) + \sum_{i=1}^{d} \delta_i \frac{\partial \pi_\theta(a|s)}{\partial \theta_i} + O(\|\delta\|^2)\,.$$

We may hope that for a richly parameterized policy class that there is a δ which moves us towards a greedy policy improvement, and, this would facilitate reaching a near optimal policy.

Define the greedy policy as follows:

$$\pi_\theta^+(s) = \text{argmax}_{a\in\mathcal{A}}\, A^{\pi_\theta}(s,a)$$

where we break ties in some predetermined order.

Motivated by this, let us now define the *Bellman policy error* in approximating π_θ^+ as follows:

$$\begin{aligned} L_{\text{BPE}}(\theta) &:= \min_{(w,w_0)} \mathbb{E}_{s\sim d^{\pi_\theta}} \left\| \pi_\theta^+(\cdot|s) - \left(w_0\pi_\theta(\cdot|s) + \sum_{i=1}^{d} w_i \frac{\partial \pi_\theta(\cdot|s)}{\partial \theta_i}\right)\right\|_1 \\ &= \min_{(w,w_0)} \mathbb{E}_{s\sim d^{\pi_\theta}} \left[\sum_a \left| \pi_\theta^+(a|s) - \left(w_0\pi_\theta(a|s) + \sum_{i=1}^{d} w_i \frac{\partial \pi_\theta(a|s)}{\partial \theta_i}\right)\right|\right] \end{aligned}$$

and define the best fit (w, w_0) as:

$$(w^*(\theta), w_0^*(\theta)) = \text{argmin}_{(w,w_0)}\, \mathbb{E}_{s\sim d^{\pi_\theta}} \left\| \pi_\theta^+(\cdot|s) - \left(w_0\pi_\theta(\cdot|s) + \sum_{i=1}^{d} w_i \frac{\partial \pi_\theta(\cdot|s)}{\partial \theta_i}\right)\right\|_1$$

Define two sources of error:

$$\begin{aligned}\epsilon_{\text{approx}}(\theta) &:= L_{\text{BPE}}(\theta)\\ \epsilon_{\text{optim}}(\theta) &:= \|\nabla V^{\pi_\theta}\|\,.\end{aligned}$$

These definitions are intended to quantify the error due to both function approximation error and optimization error.

Theorem 4.11. *(Approximate Optimality) For all θ, we have that:*

$$V^{\pi_\theta} \geq V^* - \left\|\frac{d^{\pi^*}}{d^{\pi_\theta}}\right\|_\infty \left(\epsilon_{optim}(\theta)\|w^*(\theta)\|_2 + \frac{1}{1-\gamma}\epsilon_{approx}(\theta)\right)\,,$$

where $\frac{d^{\pi^*}}{d^{\pi_\theta}}$ represents componentwise division.

Proof: To simplify notation let $(w, w_0) = (w^*(\theta), w_0^*(\theta))$. Define:

$$\hat{\pi}_\theta(a|s) = w_0\pi_\theta(a|s) + \sum_{i=1}^d w_i \frac{\partial \pi_\theta(a|s)}{\partial \theta_i}\,.$$

By the performance difference lemma (Lemma 4.2),

$$\begin{aligned}V^* - V^{\pi_\theta} &= \frac{1}{1-\gamma}\sum_{s,a} d^{\pi^*}(s)\pi^*(a|s)A^{\pi_\theta}(s,a)\\
&\leq \frac{1}{1-\gamma}\sum_{s} d^{\pi^*}(s)\max_{a\in\mathcal{A}} A^{\pi_\theta}(s,a)\\
&\leq \frac{1}{1-\gamma}\left\|\frac{d^{\pi^*}}{d^{\pi_\theta}}\right\|_\infty \sum_{s} d^{\pi_\theta}(s)\max_{a\in\mathcal{A}} A^{\pi_\theta}(s,a)\,.\\
&= \frac{1}{1-\gamma}\left\|\frac{d^{\pi^*}}{d^{\pi_\theta}}\right\|_\infty \sum_{s,a} d^{\pi_\theta}(s)\pi_\theta^+(a|s)A^{\pi_\theta}(s,a)\\
&= \frac{1}{1-\gamma}\left\|\frac{d^{\pi^*}}{d^{\pi_\theta}}\right\|_\infty \left(\sum_{s,a} d^{\pi_\theta}(s)\hat{\pi}_\theta(a|s)A^{\pi_\theta}(s,a) + \sum_{s,a} d^{\pi_\theta}(s)\Big(\pi_\theta^+(a|s) - \hat{\pi}_\theta(a|s)\Big)A^{\pi_\theta}(s,a)\right)\,.\end{aligned}$$

Now we bound each of these two terms separately. For the first term,

$$\begin{aligned}\sum_{s,a} d^{\pi_\theta}(s)\hat{\pi}_\theta(a|s)A^{\pi_\theta}(s,a) &= w_0\sum_{s,a} d^{\pi_\theta}(s)\pi_\theta(a|s)A^{\pi_\theta}(s,a) + \sum_{s,a} d^{\pi_\theta}(s)\left(\sum_{i=1}^d w_i\frac{\partial \pi_\theta(a|s)}{\partial\theta_i}\right)A^{\pi_\theta}(s,a)\\
&= 0 + \sum_{i=1}^d w_i \sum_{s,a} d^{\pi_\theta}(s)\frac{\partial \pi_\theta(a|s)}{\partial\theta_i}A^{\pi_\theta}(s,a)\\
&= (1-\gamma)w^\top \nabla V^{\pi_\theta}\\
&\leq (1-\gamma)\|w^\top\|_2\|\nabla V^{\pi_\theta}\|_2\end{aligned}$$

For the second term,

$$\begin{aligned}\sum_{s,a} d^{\pi_\theta}(s)\Big(\pi_\theta^+(a|s) - \hat{\pi}_\theta(a|s)\Big)A^{\pi_\theta}(s,a) &\leq \sum_{s,a} d^{\pi_\theta}(s)|\pi_\theta^+(a|s) - \hat{\pi}_\theta(a|s)||A^{\pi_\theta}(s,a)|\\
&\leq \sum_{s,a} d^{\pi_\theta}(s)|\pi_\theta^+(a|s) - \hat{\pi}_\theta(a|s)|\\
&= L_{\text{BPE}}(\theta)\,.\end{aligned}$$

The proof is completed by substitution. ■

4.4 Compatible Function Approximation and Preconditioning

Compatible function approximation: Following Sutton et al. [1999] , define an error function for approximating $A^{\pi_\theta}(s,a)$ as follows:

$$L^\theta(w) := \mathbb{E}_{s\sim d^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\Big(w^\top\nabla\log\pi_\theta(a|s) - A^{\pi_\theta}(s,a)\Big)^2\right].$$

Let $w_\star^\theta$ be the maximizer of $L^\theta(w)$. Define:

$$\mathcal{F}^\theta := \nabla^2 L^\theta = \mathbb{E}_{s\sim d^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[(\nabla\log\pi_\theta(a|s))\nabla\log\pi_\theta(a|s)^\top\right].$$

(note that the Hessian is not a function of w). Note that $\mathcal{F}^\theta$ is the (average) Fisher information matrix on the family of distributions $\{\pi_\theta(\cdot|s)|s\in\mathcal{S}\}$; in the next section, we see consider the use of $\mathcal{F}^\theta$ as a pre-conditioner (in the natural policy gradient algorithm [Kakade, 2001]).

The first order optimality condition for $w_\star^\theta$ is:

$$\mathbb{E}_{s\sim d^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\nabla\log\pi_\theta(a|s)\Big((\nabla\log\pi_\theta(a|s))^\top w_\star^\theta - A^{\pi_\theta}(s,a)\Big)\right] = 0 \tag{4.3}$$

whose solution is given by

$$w_\star^\theta = (1-\gamma)\big(\mathcal{F}^\theta\big)^+\nabla V^{\pi_\theta}. \tag{4.4}$$

where M^+ denotes the pseudo-inverse of M.

Sutton et al. [1999] provides the following observation:

Lemma 4.12. *(Compatible Function Approximation), Define:*

$$\widetilde{A^{\pi_\theta}}(s,a) = (w_\star^\theta)^\top\nabla\log\pi_\theta(a|s).$$

We have that:

$$\nabla V^{\pi_\theta} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\widetilde{A^{\pi_\theta}}(s,a)\nabla\log\pi_\theta(a|s)\right]$$

Proof: The first order optimality condition, Equation 4.3, for $w_\star^\theta$ may be expressed as:

$$\mathbb{E}_{s\sim d^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\nabla\log\pi_\theta(a|s)\Big(\widetilde{A^{\pi_\theta}}(s,a) - A^{\pi_\theta}(s,a)\Big)\right] = 0.$$

The proof is completed after rearranging. ■

This lemma shows that in order to accurately estimate the policy gradient we only need accurate estimates of $\widetilde{A^{\pi_\theta}}$ instead of A^{π_θ} itself. This is a helpful observation when using actor critic methods.

Preconditioning: The Natural Policy Gradient The matrix $\mathcal{F}^\theta$ can be viewed as the (average) Fisher information matrix on the family of distributions $\{\pi_\theta(\cdot|s)|s\in\mathcal{S}\}$, where the average is under the state-action visitation frequencies.

The natural policy gradient method [Kakade, 2001] is defined as:

$$\theta \leftarrow \theta + \eta(\mathcal{F}^\theta)^+\nabla V^{\pi_\theta}$$

While arguments for this method have been provided based on information geometry [Kakade, 2001, Bagnell and Schneider, 2003], it is natural to gain intuition for this preconditioner in the case of the Gibbs policy, where it has close relationship to the policy iteration algorithm.

Lemma 4.13. *(The natural gradient and policy iteration) For the Gibbs policy class, we have that:*

$$(\mathcal{F}^\theta)^+ \nabla V^{\pi_\theta} = \frac{1}{1-\gamma} A^{\pi_\theta}$$

This implies that in the limit $\eta_t \to \infty$, then:

$$\pi_{\theta_{t+1}}(a|s) = \mathrm{argmax}_a \, Q^{\theta_t}(s,a)$$

Proof: By Equation 4.4,

$$\left(\mathcal{F}^\theta\right)^+ \nabla V^{\pi_\theta} = \frac{1}{1-\gamma} w_\star^\theta .$$

Now let us show that, for the Gibbs policy, $(w_\star^\theta)_{s,a} = A^{\pi_\theta}(s,a)$, which would complete the proof. To see this, observe that

$$(A^{\pi_\theta})^\top \nabla \log \pi_\theta(a|s) = A^{\pi_\theta}(s,a) - \sum_{a'} \pi_\theta(a'|s) A^{\pi_\theta}(s,a') = A^{\pi_\theta}(s,a) .$$

due to that $\sum_{a'} \pi_\theta(a'|s) A^{\pi_\theta}(s,a') = 0$. Hence,

$$L^\theta(A^{\pi_\theta}) = \mathbb{E}_{s\sim d^{\pi_\theta}} \mathbb{E}_{a\sim \pi_\theta(\cdot|s)} \left[\Big((A^{\pi_\theta})^\top \nabla \log \pi_\theta(a|s) - A^{\pi_\theta}(s,a) \Big)^2 \right] = 0 ,$$

which implies $w_\star^\theta = A^{\pi_\theta}$ is a minimizer. ■

Chapter 5

Value Function Approximation

Reinforcement Learning and Bandits **Spring 2019**

Value Function Approximation

Instructors: Alekh Agarwal, Sham Kakade *Lecture 5*

The previous lecture showed how policy iteration like methods can be generalized to large state spaces (think discrete, but exponentially large in the desired accuracy parameters) and for policy classes that are not necessarily tabular. This chapter will consider a similar question, but for the class of value iteration like methods. We will attack the question in two parts. First we will study how the value function of a given policy can be evaluated in large state spaces, before moving to the question of policy improvement. The result and proofs in this chapter have been obtained in collaboration with Nan Jiang.

5.1 Approximate Policy Evaluation

First we need some notational preliminaries. Throughout this lecture, we will consider the problem of approximating Q-value functions. We will assume that we are given a class of candidate value functions $\mathcal{F}$, where each $f \in \mathcal{F}$ is a mapping from $\mathcal{S} \times \mathcal{A}$ to $[0, 1]$ so that it satisfies the desired semantics of a Q-value function. Given a policy π, one question of interest is to find a function $f \in \mathcal{F}$ which approximates Q^π. In order to define the notion of approximation, we introduce additional notation. Note that in our previous lectures, we typically considered approximating value functions uniformly, that is under an ℓ_∞ norm over states and actions. While this is feasible in the tabular case, once the number of states grows large, we can no longer hope to find such a function f which will approximate Q^π uniformly for all states and actions from a reasonable amount of data, and with a function class $\mathcal{F}$ of reasonable statistical complexity (these concepts will be made precise in the sequel). In order to sidestep this issue, taking a cue from supervised learning, we will instead switch to measuring our errors in expectation over some state distribution.

Concretely, given a function $g : \mathcal{S} \times \mathcal{A} \to [0, 1]$, and a distribution μ over $\mathcal{S} \times \mathcal{A}$, we define:

$$\|g\|_\mu := \sqrt{\mathbb{E}_{(s,a)\sim\mu} g^2(x)}. \tag{5.1}$$

For instance, if $g(s, a) = f(s, a) - Q^\pi(s, a)$, then $\|g\|_\mu$ measures the expected squared loss of f in approximating Q^π under the distribution μ over $\mathcal{S}$ and $\mathcal{A}$. For evaluating the quality of f in approximating Q^π, a natural measure to consider is

$$d^\pi(s)\pi(a \mid s) = d^\pi_{d_0}(s)\pi(a \mid s) = \mathbb{E}_{s_0\sim d_0}[(1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi(s_t = s|s_0)]\pi(a \mid s),$$

where the notation $d^\pi_{d_0}(s)$ is consistent with our definition from the previous lecture. As before, we will drop the subscript d_0 as it will be fixed throughout our treatment. We will further overload the $\|g\|_\mu$ notation, and also allow the use of unnormalized and signed measures μ, interpreting them as integrals with respect to an appropriate base measure. We quickly visit a few useful properties of this generalization, the proofs of which are left to the reader.

Fact 5.1. The definition (5.1) of $\|g\|_\mu$, once extended to signed and unnormalized measures implies that:

1. $\|g\|_{c\mu} = \sqrt{c}\,\|g\|_\mu$.
2. $\|g\|_{\mu_1-\mu_2} \leq \|g\|_{\mu_1}$, whenever $\|g\|^2_{\mu_1-\mu_2} \geq 0$.

We further define the Bellman operator $\mathcal{T}^\pi$. For a function $g : \mathcal{S} \times \mathcal{A} \to [0, 1]$, we have

$$\mathcal{T}^\pi[g](s,a) := (1-\gamma)r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)} g(s', \pi(s')). \tag{5.2}$$

We now define a simple procedure for estimating the function Q^π. Suppose we are given a set of N samples (s, a, s') drawn according to $d^\pi(s)\nu(a \mid s) \times P(s'|s,a)$. Here $\nu(a|s)$ is an arbitrary distribution used for generating actions. While π might be a natural choice of ν as mentioned earlier, we allow more general choices as the question of policy improvement which we consider next will require our Q-value estimates to be precise not just for the actions chosen by π, as we will see in the sequel. A reasonable mental model for generating such samples is to generate N trajectories, where we take a random number of steps according to π, starting from the initial state drawn according to d_0, and then take one more action as per ν. It is easily seen that this sampling model induces the desired probability distribution over the (s, a, s') triples. We will use the shorthand $d^{\pi,\nu}$ to denote the joint distribution over (s, a) pairs which draws $s \sim d^\pi_{d_0}$ and $a \sim \nu(\cdot \mid s)$.

Let (s_i, a_i, s'_i) refer to the i_{th} triple for $i = 1, 2, \ldots, N$. Let $s_{t,i}$ denote the state encountered at time t in trajectory i and similarly for the actions. Let us define:

$$\begin{aligned}
\mathcal{L}(f; f') &= \mathbb{E}_{(s,a)\sim d^{\pi,\nu}, s'\sim P(\cdot|s,a)}[(f(s,a) - (1-\gamma)r(s,a) - \gamma f'(s', \pi(s')))^2].\\
\widehat{\mathcal{L}}_N(f; f') &= \frac{1}{N}\sum_{i=1}^N (f(s_i,a_i) - (1-\gamma)r(s_i,a_i) - \gamma f'(s'_i, \pi(s'_i)))^2.\\
\widehat{\mathcal{T}}^\pi_{\mathcal{F}} f' &= \mathrm{argmin}_{f\in\mathcal{F}}\, \widehat{\mathcal{L}}_N(f; f').
\end{aligned}$$

Intuitively, $\widehat{\mathcal{T}}^\pi_{\mathcal{F}}(f')$ maps a function f' to an f which minimizes the Bellman error on π's trajectories, when the future values are predicted under f'. For the population loss $\mathcal{L}(f; f')$, we expect it to be minimized when $f = f' = Q^\pi$ due to the Bellman equations for value functions.

The procedure we consider for obtaining an estimate of Q^π from samples is the following sample-based value iteration. Initialize $f_0 \in \mathcal{F}$ arbitrarily and iterate:

$$f_k = \widehat{\mathcal{T}}^\pi_{\mathcal{F}} f_{k-1} \quad \text{for } k > 0. \tag{5.3}$$

For this procedure, we will establish the following convergence guarantee.

Theorem 5.2. *For the procedure defined in Equation 5.3, for all $k = 0, 1, 2, \ldots$, with probability at least $1-\delta$,*

$$\|f_k - Q^\pi\|_{d^{\pi,\nu}} \le \sqrt{\|\rho\|_\infty \|1/\rho\|_\infty}\gamma^{k/2}\|f_0 - Q^\pi\|_{d^{\pi,\nu}} + \frac{2\sqrt{\|\rho\|_\infty \|1/\rho\|_\infty}}{1-\gamma}\left(\sqrt{3}\max_{f\in\mathcal{F}}\|\mathcal{T}^\pi_{\mathcal{F}} f - \mathcal{T}^\pi f\|_{d^{\pi,\nu}} + \sqrt{\frac{19}{N}\log\frac{|\mathcal{F}|}{\delta}}\right).$$

While the second part of the error bound which depends on N goes to zero as we increase the number of samples, the term $\max_{f\in\mathcal{F}}\|\mathcal{T}^\pi\mathcal{F} f - \mathcal{T}^\pi f\|_{d^\pi}$ does not asymptotically vanish in general. It is easily seen to be 0 in the tabular function class, but can be arbitrarily large for other classes. This quantity has been referred to as the *inherent one-step Bellman error* in Antos et al. [2008] and is well-known to control the quality of function approximation guarantees.

We will analyze the properties of this procedure in two parts. First we will obtain an error bound on the quality of f_k in approximating Q^π, assuming all the sample-based estimates are close to their expectations. We will then quantify the size of the statistical deviations.

The error measure we will use to study the convergence of f_k to Q^π is $\|f_k - Q\|_{d^\pi}$. We start with a simple lemma.

Lemma 5.3 (Error decomposition)**.** *For any sequence $f_k : \mathcal{S} \times \mathcal{A} \to [0,1]$, we have the bound*

$$\|f_k - Q^\pi\|_{d^\pi \nu} \leq \gamma^{k/2} \sqrt{\|\rho\|_\infty \|1/\rho\|_\infty}\, \|f_0 - Q^\pi\|_{d^{\pi,\nu}} + \sum_{i=1}^{k} \gamma^{(k-i)/2} \sqrt{\|\rho\|_\infty \|1/\rho\|_\infty}\, \|f_i - \mathcal{T}^\pi f_{i-1}\|_{d^{\pi,\nu}}$$

Proof: We will prove the lemma in two steps. First we establish a bound on $\|f_k - Q^\pi\|_{d^{\pi,\pi}}$ in terms of $\|f_i - \mathcal{T}^\pi f_{i-1}\|_{d^{\pi,\pi}}$. We then apply importance weighting to correct for the measure mismatch on both sides of the inequality to obtain the final bound.

For any scaled probability measure μ, d^μ is always a norm so that it satisfies triangle inequality. This implies

$$\begin{aligned}
\|f_k - Q^\pi\|_{d^{\pi,\pi}} &\leq \|f_k - \mathcal{T}^\pi f_{k-1}\|_{d^{\pi,\pi}} + \|\mathcal{T}^\pi f_{k-1} - Q^\pi\|_{d^{\pi,\pi}} \\
&= \|f_k - \mathcal{T}^\pi f_{k-1}\|_{d^{\pi,\pi}} + \|\mathcal{T}^\pi f_{k-1} - \mathcal{T}^\pi Q^\pi\|_{d^{\pi,\pi}} && \text{(Bellman equation for } Q^\pi) \\
&= \|f_k - \mathcal{T}^\pi f_{k-1}\|_{d^{\pi,\pi}} + \gamma \|P^\pi f_{k-1} - P^\pi Q^\pi\|_{d^{\pi,\pi}} \\
&\leq \|f_k - \mathcal{T}^\pi f_{k-1}\|_{d^{\pi,\pi}} + \gamma \|f_{k-1} - Q^\pi\|_{d^{\pi,\pi} P^\pi},
\end{aligned}$$

where the final equality treats $d^{\pi,\pi}$ as a row vector with cardinality $|\mathcal{S}| \cdot |\mathcal{A}|$. This equality follows from the definition of $\mathcal{T}^\pi$ and since the reward function is assumed to be known. The last inequality is a consequence of Jensen's inequality which gives that $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$. Note that by the first part of Fact 5.1, we further obtain

$$\|f_k - Q^\pi\|_{d^{\pi,\pi}} \leq \|f_k - \mathcal{T}^\pi f_{k-1}\|_{d^{\pi,\pi}} + \sqrt{\gamma}\, \|f_{k-1} - Q^\pi\|_{\gamma d^{\pi,\pi} P^\pi}.$$

In order to simplify further, we define $\mu_0(s,a) = d_0(s)\pi(a \mid s)$ to be the initial measure extended to (s,a) pairs and observe that $d^{\pi,\pi}$ can be written as

$$d^{\pi,\pi} = (1-\gamma)\mu_0 (I - \gamma P^\pi)^{-1} \quad \text{so that} \quad \gamma P^\pi d^{\pi,\pi} = d^{\pi,\pi} - (1-\gamma)\mu_0,$$

where the inequalities hold elementwise. Substituting this in our earlier bound, we further obtain

$$\begin{aligned}
\|f_k - Q^\pi\|_{d^{\pi,\pi}} &\leq \|f_k - \mathcal{T}^\pi f_{k-1}\|_{d^{\pi,\pi}} + \sqrt{\gamma}\, \|f_{k-1} - Q^\pi\|_{d^{\pi,\pi} - (1-\gamma)\mu_0} \\
&\leq \|f_k - \mathcal{T}^\pi f_{k-1}\|_{d^{\pi,\pi}} + \sqrt{\gamma}\, \|f_{k-1} - Q^\pi\|_{d^{\pi,\pi}} && \text{(Part 2 of Fact 5.1)} \\
&\leq \gamma^{k/2} \|f_0 - Q^\pi\|_{d^{\pi,\pi}} + \sum_{i=1}^{k} \gamma^{(k-i)/2} \|f_i - \mathcal{T}^\pi f_{i-1}\|_{d^{\pi,\pi}},
\end{aligned}$$

where we obtain the final inequality by iterating the same set of steps to bound $\|f_i - Q^\pi\|_{d^\pi}$ for each $i = 1, 2, \ldots, k$.

In order to complete the proof, let us recall the notation $\rho(s,a) = \nu(s,a)/\pi(s,a)$. Note that for any two functions $f_1, f_2 \in \{\mathcal{S} \times \mathcal{A} \to [0,1]\}$, $\|f_1\|_{d^{\pi,\pi}} \leq \|f_2\|_{d^{\pi,\pi}}$ implies that

$$\|f_1\|_{d^{\pi,\nu}} \leq \|f_1\|_{\|\rho\|_\infty d^{\pi,\pi}} \leq \|f_2\|_{\|\rho\|_\infty d^{\pi,\pi}} \leq \|f_1\|_{\|\rho\|_\infty \|1/\rho\|_\infty d^{\pi,\nu}}.$$

Combining this with our earlier bound completes the proof of the lemma. ∎

The bound above suggests that if we had good control over the errors $\|f_i - \mathcal{T}^\pi f_{i-1}\|$ at each iteration, then we get successively good approximations to Q^π. Our next lemma takes a step in this direction by relating the loss function $\mathcal{L}$ whose empirical counterpart is minimized in our iteration (5.3) and the Bellman errors which show up in the previous bound.

Lemma 5.4. *For any $f' \in \mathcal{F}$, we have*

$$\|f - \mathcal{T}^\pi f'\|^2_{d^{\pi,\nu}} = \mathcal{L}(f; f') - \mathcal{L}(\mathcal{T}^\pi f'; f').$$

Proof: The proof essentially follows using standard properties of least squares regression. Note that

$$\begin{aligned}
\mathcal{L}(f; f') &= \mathbb{E}_{(s,a)\sim d^{\pi,\nu}, s'\sim P(\cdot|s,a)}[(f(s,a) - (1-\gamma)r(s,a) - \gamma f'(s', \pi(s')))^2] \\
&= \mathbb{E}_{(s,a)\sim d^{\pi,\nu}, s'\sim P(\cdot|s,a)}[(f(s,a) - \mathcal{T}^\pi f'(s,a) + \mathcal{T}^\pi f'(s,a) - (1-\gamma)r(s,a) - \gamma f'(s', \pi(s')))^2] \\
&= \|f - \mathcal{T}^\pi f'\|^2_{d^\pi} + \mathcal{L}(\mathcal{T}^\pi f'; f') \\
&\quad + 2\mathbb{E}_{(s,a)\sim d^{\pi,\nu}, s'\sim P(\cdot|s,a)}(f(s,a) - \mathcal{T}^\pi f'(s,a))(\mathcal{T}^\pi f'(s,a) - (1-\gamma)r(s,a) - \gamma f'(s', \pi(s'))) \\
&= \|f - \mathcal{T}^\pi f'\|^2_{d^{\pi,\nu}} + \mathcal{L}(\mathcal{T}^\pi f'; f') + 2\mathbb{E}\left[(f(s,a) - \mathcal{T}^\pi f'(s,a))\mathbb{E}[\mathcal{T}^\pi f'(s,a) - (1-\gamma)r(s,a) - \gamma f'(s', \pi(s')) \mid s,a]\right] \\
&= \|f - \mathcal{T}^\pi f'\|^2_{d^{\pi,\nu}} + \mathcal{L}(\mathcal{T}^\pi f'; f'). \qquad \text{(since } \mathbb{E}[(1-\gamma)r(s,a) - \gamma f'(s', \pi(s')) \mid s,a] = \mathcal{T}^\pi f'(s,a))
\end{aligned}$$

■

This lemma has a particularly useful corollary for approximate minimizers of the loss. For stating the result, let us define the notation $\mathcal{T}^\pi_{\mathcal{F}} f' = \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{L}(f; f')$ and the corresponding empirical counterparts as well. Then we have the following result.

Corollary 5.5. *If f is an ϵ suboptimal minimizer of $\mathcal{L}$ over $\mathcal{F}$, that is, $\mathcal{L}(f; f') \le \mathcal{L}(\mathcal{T}^\pi_{\mathcal{F}} f'; f) + \epsilon$, then $\|f - \mathcal{T}^\pi f'\|^2_{d^{\pi,\nu}} \le \|\mathcal{T}^\pi_{\mathcal{F}} f' - \mathcal{T}^\pi f'\|^2_{d^{\pi,\nu}} + \epsilon$.*

Proof: The ϵ-optimality of f implies that

$$\begin{aligned}
\|f - \mathcal{T}^\pi f'\|^2_{d^{\pi,\nu}} &= \mathcal{L}(f; f') - \mathcal{L}(\mathcal{T}^\pi f'; f') \\
&\le \mathcal{L}(\mathcal{T}^\pi_{\mathcal{F}} f'; f') + \epsilon - \mathcal{L}(\mathcal{T}^\pi f'; f') \\
&= \|\mathcal{T}^\pi_{\mathcal{F}} f' - \mathcal{T}^\pi f'\|^2_{d^{\pi,\nu}} + \epsilon,
\end{aligned}$$

where the first inequality uses the approximate optimality of f and $\mathcal{T}^\pi_{\mathcal{F}} f' \in \mathcal{F}$ while the second uses Lemma 5.4 with $f = \mathcal{T}^\pi_{\mathcal{F}} f'$. ■

The corollary has a rather intuitive interpretation. If f approximately minimizes $\mathcal{L}$, then it is also close to $\mathcal{T}^\pi f'$ up to an amount which depends on the suboptimality as well as the distance of $\mathcal{T}^\pi f'$ to $\mathcal{F}$. The second term can be thought of as an approximation error, and is always zero, for example, when the function class is fully tabular.

Since f_k minimizes the empirical loss $\widehat{\mathcal{L}}(f; f_{k-1})$, we expect it to approximately minimize the population version $\mathcal{L}(f; f_{k-1})$ by invoking concentration arguments. If we can use this intuition to quantify ϵ, then we can plug the result of Corollary 5.5 into that of Lemma 5.3 to quantify the convergence of f_i to Q^π. We give a lemma that gives a bound on ϵ next. We give the simplest version of this argument assuming that the N samples we use to construct $\widehat{\mathcal{L}}$ are drawn i.i.d., but similar arguments continue to hold if the samples are drawn according to a suitably regular Markov chain, as is the case when they are consecutive triples from a trajectory. For the ease of presentation, we will also restrict $\mathcal{F}$ to a large, but finite class so that we can control statistical deviations uniformly over all functions in $\mathcal{F}$ using a union bound.

Lemma 5.6. *For the sequence f_k, $k = 0, 1, \ldots, \infty$ as defined in Equation 5.3, with probability at least $1-\delta$, we have*

$$\mathcal{L}(f_k; f_{k-1}) - \mathcal{L}(\mathcal{T}^\pi_{\mathcal{F}} f_{k-1}; f_{k-1}) \le 2\,\|\mathcal{T}^\pi_{\mathcal{F}} f_{k-1} - \mathcal{T}^\pi f_{k-1}\|^2_{d^{\pi,\nu}} + \frac{19}{N}\log\frac{|\mathcal{F}|}{\delta}.$$

Proof: For $i = 1, 2, \dots, N$, and given two functions f, f', let us define

$$X_i = (f(s_i, a_i) - (1-\gamma)r(s_i, a_i) - \gamma f'(s_i', \pi(s_i')))^2 - (\mathcal{T}_\mathcal{F} f'(s_i, a_i) - (1-\gamma)r(s_i, a_i) - \gamma f'(s_i', \pi(s_i')))^2.$$

It is easily checked that

$$\widehat{\mathcal{L}}(f; f') - \widehat{\mathcal{L}}(\mathcal{T}_\mathcal{F} f'; f') = \frac{1}{N}\sum_{i=1}^N X_i.$$

Furthermore, the random variables X_i are i.i.d. by assumption, bounded by 1 in absolute value and the variance is bounded by

$$\begin{aligned}
\mathbb{E}[X_i^2] &= \mathbb{E}[(f(s_i, a_i) - \mathcal{T}_\mathcal{F}^\pi f'(s_i, a_i))^2 (f(s_i, a_i) + \mathcal{T}_\mathcal{F}^\pi f'(s_i, a_i) - 2(1-\gamma)r(s_i, a_i) - 2\gamma f'(s_i', \pi(s_i')))^2] \\
&\le 4\mathbb{E}[(f(s_i, a_i) - \mathcal{T}_\mathcal{F}^\pi f'(s_i, a_i))^2] = 4\left\|f - \mathcal{T}_\mathcal{F}^\pi f'\right\|_{d^{\pi,\nu}}^2 \\
&\le 8(\|f - \mathcal{T}^\pi f'\|_{d^{\pi,\nu}}^2 + \|\mathcal{T}_\mathcal{F}^\pi f' - \mathcal{T}^\pi f'\|_{d^{\pi,\nu}}^2) && \text{(Cauchy-Schwarz inequality)} \\
&= 8(\mathcal{L}(f; f') - \mathcal{L}(\mathcal{T}^\pi f'; f') + \|\mathcal{T}_\mathcal{F}^\pi f' - \mathcal{T}^\pi f'\|_{d^{\pi,\nu}}^2) && \text{(Lemma 5.4)} \\
&= 8(\mathcal{L}(f; f') - \mathcal{L}(\mathcal{T}_\mathcal{F}^\pi f'; f') + 2\left\|\mathcal{T}_\mathcal{F}^\pi f' - \mathcal{T}^\pi f'\right\|_{d^{\pi,\nu}}^2). && \text{(Lemma 5.4 with } f = \mathcal{T}_\mathcal{F}^\pi f')
\end{aligned}$$

Since $\mathbb{E}[X_i] = \mathcal{L}(f; f') - \mathcal{L}(\mathcal{T}_\mathcal{F}^\pi f'; f') \ge 0$, invoking Bernstein's inequality yields that with probability $1-\delta$, simultaneously for all pairs $f, f' \in \mathcal{F}$

$$\begin{aligned}
&\mathcal{L}(f; f') - \mathcal{L}(\mathcal{T}_\mathcal{F}^\pi f'; f') - \widehat{\mathcal{L}}(f; f') + \widehat{\mathcal{L}}(\mathcal{T}_\mathcal{F}^\pi f'; f') \\
&\le \sqrt{\frac{16}{N}(\mathcal{L}(f; f') - \mathcal{L}(\mathcal{T}_\mathcal{F}^\pi f'; f') + 2\left\|\mathcal{T}_\mathcal{F}^\pi f' - \mathcal{T}^\pi f'\right\|_{d^{\pi,\nu}}^2)\log\frac{|\mathcal{F}|^2}{\delta}} + \frac{2}{3N}\log\frac{|\mathcal{F}|^2}{\delta} \\
&\le \frac{1}{2}(\mathcal{L}(f; f') - \mathcal{L}(\mathcal{T}_\mathcal{F}^\pi f'; f') + 2\left\|\mathcal{T}_\mathcal{F}^\pi f' - \mathcal{T}^\pi f'\right\|_{d^{\pi,\nu}}^2) + \frac{8}{N}\log\frac{|\mathcal{F}|}{\delta} + \frac{4}{3N}\log\frac{|\mathcal{F}|}{\delta},
\end{aligned}$$

where the second inequality uses $\sqrt{ab} \le (a+b)/2$ from Cauchy-Schwarz inequality. Rearranging terms, we see that with probability at least $1-\delta$

$$\mathcal{L}(f; f') - \mathcal{L}(\mathcal{T}_\mathcal{F}^\pi f'; f') \le 2(\widehat{\mathcal{L}}(f; f') - \widehat{\mathcal{L}}(\mathcal{T}_\mathcal{F}^\pi f'; f')) + 2\left\|\mathcal{T}_\mathcal{F}^\pi f' - \mathcal{T} f'\right\|_{d^{\pi,\nu}}^2 + \frac{19}{N}\log\frac{|\mathcal{F}|}{\delta}.$$

Choosing $f' = f_{k-1}$ and $f = f_k$ so that f_k minimizes $\widehat{\mathcal{L}}(f; f_{k-1})$ in the inequality above, we see that with probability at least $1-\delta$,

$$\mathcal{L}(f_k; f_{k-1}) - \mathcal{L}(\mathcal{T}_\mathcal{F}^\pi f_{k-1}; f_{k-1}) \le 2\left\|\mathcal{T}_\mathcal{F}^\pi f_{k-1} - \mathcal{T}^\pi f_{k-1}\right\|_{d^{\pi,\nu}}^2 + \frac{19}{N}\log\frac{|\mathcal{F}|}{\delta}.$$

In order to obtain a uniform bound across all iterations, we note that the bound here holds for all pairs f, f' so that across all the iterations, the failure probability is at most δ. ■

Having obtained a bound on the statistical errors, we are now ready to prove the theorem.

Proof:[Proof of Theorem 5.2] The proof largely combines the results of our lemmas so far. We condition on the high probability event in Lemma 5.6 since it hods with probability at least $1-\delta$ throughout the run of the algorithm. Under this event, combining Lemma 5.6 along with Corollary 5.5 yields that

$$\|f_k - \mathcal{T}^\pi f_{k-1}\|_{d^{\pi,\nu}}^2 \le 3\left\|\mathcal{T}_\mathcal{F}^\pi f_{k-1} - \mathcal{T}^\pi f_{k-1}\right\|_{d^{\pi,\nu}}^2 + \frac{19}{N}\log\frac{|\mathcal{F}|}{\delta}$$

Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we see that for all $i = 1, 2, \dots, k$

$$\|f_i - \mathcal{T}^\pi f_{i-1}\|_{d^{\pi,\nu}} \le \sqrt{3}\max_{f\in\mathcal{F}}\|\mathcal{T}_\mathcal{F}^\pi f - \mathcal{T}^\pi f\|_{d^{\pi,\nu}} + \sqrt{\frac{19}{N}\log\frac{|\mathcal{F}|}{\delta}} := \phi.$$

Plugging this result into the conclusion of Lemma 5.3, we obtain

$$\begin{aligned}\|f_k - Q^\pi\|_{d^{\pi,\nu}} &\leq \sqrt{\|\rho\|_\infty \|1/\rho\|_\infty}\gamma^{k/2}\|f_0 - Q^\pi\|_{d^{\pi,\nu}} + \sqrt{\|\rho\|_\infty \|1/\rho\|_\infty}\sum_{i=1}^{k}\gamma^{(k-i)/2}\phi \\ &\leq \sqrt{\|\rho\|_\infty \|1/\rho\|_\infty}\left(\gamma^{k/2}\|f_0 - Q^\pi\|_{d^{\pi,\nu}} + \frac{\phi}{1-\sqrt{\gamma}}\right) \leq \sqrt{\|\rho\|_\infty \|1/\rho\|_\infty}\left(\gamma^{k/2}\|f_0 - Q^\pi\|_{d^{\pi,\nu}} + \frac{2\phi}{1-\gamma}\right).\end{aligned}$$

Recalling the value of ϕ completes the proof of the theorem. ∎

5.2 Approximate Policy Improvement

Evaluating the function Q^π for a given policy π is often an intermediate step in the larger algorithm, and we now visit the question of how we might leverage the policy evaluation result of the previous section. We start with a basic result, which resembles classical bounds for approximate policy iteration, but has some differences due to our use of Q functions rather than V functions. These bounds are apt for tabular function classes and small state spaces, as they rely on a uniformly good approximation to Q^π at all states and actions. After seeing the underlying reasons that give rise to such limitations, we will consider a better policy improvement operator described in the Conservative Policy Iteration [Kakade and Langford, 2002], and discuss using the results of our previous section as a plug-in to that algorithm.

5.2.1 Greedy policy improvement with ℓ_∞ approximation

The setting of this section closely resembles that of the Policy Iteration algorithm in Chapter 1.2. We consider the following iterative procedure, which starts with an initial policy π_0 and updates for $i = 0, 1, 2, \ldots$:

$$\text{Obtain } \widehat{Q}_i \text{ such that } \|\widehat{Q}_i - Q^{\pi_i}\|_\infty \leq \epsilon. \text{ Set } \pi_{i+1}(s) = \mathrm{argmax}_{a\in\mathcal{A}}\, \widehat{Q}_i(s,a) \text{ for all } s \in \mathcal{S}. \tag{5.4}$$

We do not discuss how to obtain such an approximation $\widehat{Q}$ for Q^{π_i} here, but instead try to understand the reason why need such a uniformly good approximation, and how we might relax the condition. Intuitively, the greedy policy improvement in Equation 5.4 can induce a policy π_{i+1} which is dramatically different from π_i in its state visitation distribution. If we only had approximation guarantees between $\widehat{Q}$ and Q^π on d^{π_i}, we can already guess that it would not be sufficient as the approximation might be arbitrarily worse on $d^{\pi_{i+1}}$, and if this happens, then π_{i+1} is being defined rather arbitrarily on the states which are not visited under d^{π_i}. That is, we are effectively encountering another instance of the distribution mismatch which we also saw in the policy gradient analysis earlier.

In order to quantify these intuitions, we will establish the following result regarding the iteration (5.4).

Theorem 5.7. *The approximate policy iteration algorithm described in Algorithm 5.4 satisfies for all $k = 0, 1, 2, \ldots$*

$$\|Q^\star - Q^{\pi_k}\|_\infty \leq \gamma^k \|Q^\star - Q^{\pi_0}\|_\infty + \frac{2\epsilon\gamma}{(1-\gamma)^2}.$$

Proof: As in the case of Theorem 1.9, the main step is to show that the distance between $Q^\star$ and Q^{π_i} contracts when we do one step of policy improvement. However, we now need to account for the approximation in $\widehat{Q}_i$ in order to show that the contraction is not offset by the errors in our estimates. We start with some simple definitions and properties. For a policy π, let $\widehat{Q}$ be a function satisfying $\|\widehat{Q} - Q^\pi\|_\infty \leq \epsilon$, and let

$$\pi^+(s) = \mathrm{argmax}_{a\in\mathcal{A}}\, Q^\pi(s,a) \quad \text{and} \quad \widetilde{\pi}(s) = \mathrm{argmax}_{a\in\mathcal{A}}\, \widehat{Q}(s,a).$$

Based on the definitions of $\widehat{Q}$ and $\widetilde{\pi}$, it is easily verified that for all states s, $A^\pi(s, \widetilde{\pi}(s)) \geq -2\epsilon$. That is, $\widetilde{\pi}(s)$ is a 2ϵ-approximate greedy action with respect to $Q^\pi(s,a)$. Based on this, the performance difference lemma (Lemma 4.2) gives that

$$\begin{aligned}
Q^{\widetilde{\pi}}(s,a) - Q^\pi(s,a) &= \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a)(V^{\widetilde{\pi}}(s') - V^\pi(s')) \\
&= \frac{gamma}{1-\gamma} \sum_{s' \in \mathcal{S}} P(s'|s,a) \sum_{s'' \in \mathcal{S}} d^{\widetilde{\pi}}_{s'} A^\pi(s'', \widetilde{\pi}(s'')) \\
&\geq -\frac{2\epsilon\gamma}{1-\gamma}.
\end{aligned}$$

In the middle equality, we recall the notation d^π_s for the distribution over states induced under π starting from the state s. Now we can lower bound the one-step policy improvement. For this, recall the notation $\mathcal{T}$ for the Bellman operator which maps $\mathcal{T}Q = (1-\gamma)r + \gamma P V_Q$, where $V_Q(s) = \max_{a \in \mathcal{A}} Q(s,a)$

$$\begin{aligned}
Q^{\widetilde{\pi}}(s,a) - \mathcal{T}Q^\pi(s,a) &= \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a)(V^{\widetilde{\pi}}(s') - \max_{a' \in \mathcal{A}} Q^\pi(s',a')) \\
&\geq \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a)(V^{\widetilde{\pi}}(s') - Q^\pi(s', \widetilde{\pi}(s')) - 2\epsilon) \\
&= -2\epsilon\gamma + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a)(Q^{\widetilde{\pi}}(s', \widetilde{\pi}(s')) - Q^\pi(s', \widetilde{\pi}(s'))) \\
&\geq -2\epsilon\gamma - \frac{2\epsilon\gamma^2}{1-\gamma} = -\frac{2\epsilon\gamma}{1-\gamma},
\end{aligned}$$

where the first inequality uses the observation that $A^\pi(s', \widetilde{\pi}(s')) \geq -2\epsilon$, while second inequality from our earlier bound on the difference between $Q^{\widetilde{\pi}}$ and Q^π.

Now following the same logic as in the convergence analysis of policy iteration, we see that

$$\begin{aligned}
\|Q^\star - Q^{\widetilde{\pi}}\|_\infty &\leq \|Q^\star - \mathcal{T}Q^\pi\|_\infty + \frac{2\epsilon\gamma}{1-\gamma} \\
&= \|\mathcal{T}Q^\star - \mathcal{T}Q^\pi\|_\infty + \frac{2\epsilon\gamma}{1-\gamma} \\
&\leq \gamma\|Q^\star - Q^\pi\|_\infty + \frac{2\epsilon\gamma}{1-\gamma}.
\end{aligned}$$

Applying this inequality between every pair of successive iterates π_i and π_{i+1} and summing the resulting geometric series yields the theorem. ■

While similar results have been shown before for approximate policy iteration where V-function estimates are used instead of the Q-function estimates (see e.g. Chapter 6.2 in [Bertsekas and Tsitsiklis, 1996]), we find this version more natural as greedy maximization using V functions requires knowledge of the MDP dynamics.

Inspecting the proof closely also show the challenges with relaxing the ℓ_∞ assumption on the closeness between $\widehat{Q}$ and Q^π. Note that the performance difference lemma invoked in the first part of the proof requires that the estimates $\widehat{Q}$ be accurate on the state distribution induced by $\widetilde{\pi}$. This is not feasible using techniques from the previous section, as $\widetilde{\pi}$ is a function of $\widehat{Q}$ itself, and could pick out a pathological state distribution. One natural intuition to fix this difficulty is by algorithmically controlling the distribution mismatch between successive policies, which we will see in our next approach.

5.2.2 Conservative Policy Iteration

As the name suggests, we will now describe a more conservative version of the policy iteration algorithm, which shifts the next policy away from the current policy with a small step size to prevent drastic shifts in successive state distributions. The algorithm and analysis is adapted from the original result of Kakade and Langford [2002].

As before, the algorithm will iteratively generate a sequence of policies π_i, and we will assume that each time we can find a Q-value estimate for π_i, denoted as $\widehat{Q}_i$ which satisfies

$$\left\| \widehat{Q}_i - Q^{\pi_i} \right\|_{d^{\pi},U} \leq \epsilon, \tag{5.5}$$

where U denotes the uniform distribution over actions. Given such a guarantee, it is easily seen that for any randomized policy ν we have

$$\left\| \widehat{Q}_i - Q^{\pi_i} \right\|_{d^{\pi},\nu} \leq \sqrt{|\mathcal{A}|}\epsilon. \tag{5.6}$$

Given access to such estimates of Q-value functions, we now describe the Conservative Policy Iteration (CPI) algorithm. For the algorithm, we use $\pi_\alpha = (1-\alpha)\pi + \alpha\pi'$ to refer to a randomized policy which chooses an action according to π with probability $1-\alpha$ and according to π' with probability α.

Algorithm 2 Conservative Policy Iteration (CPI)

Input: Initial policy π_0, accuracy parameter ϵ.

1: **for** $i = 0, 1, 2 \ldots$ **do**

2: Obtain a Q value approximation $\widehat{Q}_i$ satisfying (5.5) with parameter $\epsilon/2\sqrt{|\mathcal{A}|}$.

3: Define $\widetilde{\pi}_i(s) = \operatorname{argmax}_{a\in\mathcal{A}} \widehat{Q}_i(s,a)$, and let $\widehat{A}_i = \mathbb{E}_{s\in d^{\pi_i}} \widehat{Q}_i(s, \widetilde{\pi}_i(s)) - \widehat{Q}_i(s, \pi_i(s))$.

4: If $\widehat{A}_i \leq 2\epsilon$, return π_i.

5: Update $\pi_{i+1}(a \mid s) = (1-\alpha_i)\pi_i(a \mid s) + \alpha_i \widetilde{\pi}_i(a \mid s)$, for $\alpha_i = (\widehat{A}_i - \epsilon)(1-\gamma)/4$.

6: **end for**

The main intuition behind the algorithm is that the stepsize α controls the difference between state distributions of π_i and π_{i+1}. In particular, with high probability, the trajectories generated from π_i and π_{i+1} differ in only one action. The following result formalizes this iteration at one step of the algorithm.

Theorem 5.8 (One step improvement in CPI)**.** *Let $\pi = \pi_i$, $\widetilde{\pi} = \widetilde{\pi}_i$ and $\nu = \pi_{i+1}$ be the successive policies at one iteration of Algorithm 2 and $A = \mathbb{E}_{s\sim d^{\pi}} A^{\pi}(s, \widetilde{\pi}(s))$. Then*

$$V^{\nu} - V^{\pi} \geq \frac{\alpha}{1-\gamma}\left(A - \frac{2\alpha\gamma}{1-\gamma(1-\alpha)}\right).$$

Proof: By performance difference lemma, we know that

$$V^{\nu} - V^{\pi} = \sum_{t=0}^{\infty} \gamma^t \sum_{s\in\mathcal{S}} \mathbb{P}^{\nu}(s_t = s) \sum_a \nu(a \mid s) A^{\pi}(s,a).$$

Now on a time t, with probability $(1-\alpha)^t$, all the actions have been chosen as per π on the previous time steps. If any actions were chosen by ν with the remaining probability $1-(1-\alpha)^t$, we can simply lower bound the advantage function by its largest value on that event. Doing so, we see that

$$\begin{aligned} V^{\nu} - V^{\pi} &\geq \sum_{t=0}^{\infty} \gamma^t \left((1-\alpha)^t \sum_{s\in\mathcal{S}} \mathbb{P}^{\pi}(s_t = s) \sum_a \nu(a \mid s) A^{\pi}(s,a) - (1-(1-\alpha)^t) \max_s \sum_a \nu(a \mid s) A^{\pi}(s,a) \right) \\ &\geq, \sum_{t=0}^{\infty} \gamma^t \left(\sum_{s\in\mathcal{S}} \mathbb{P}^{\pi}(s_t = s) \sum_a \nu(a \mid s) A^{\pi}(s,a) - 2(1-(1-\alpha)^t) \max_s \sum_a \nu(a \mid s) A^{\pi}(s,a) \right). \end{aligned}$$

By definition of ν, for any state s, we observe that $\sum_a \nu(a \,|\, s) A^\pi(s,a) = \alpha A^\pi(s, \widetilde{\pi}(s)) \leq \alpha$. Summing the geometric series, we see that

$$V^\nu - V^\pi \geq \frac{\alpha}{1-\gamma}\mathbb{E}_{s\sim d^\pi} A^\pi(s, \widetilde{\pi}(s)) - 2\alpha\left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha)}\right).$$

Simplifying terms yields the result. ■

A crucial aspect of this theorem is that it always requires a good estimate of the advantage function of π only under d^π, which is the main upshot of the conservative updates. Our next result bounds the gap between the quantities $\widehat{A}_i$ and the true advantages of $\widetilde{\pi}_i$ over π_i at each iteration.

Lemma 5.9. *Using the notation of Theorem 5.8, under the conditions of Equation 5.6, we have* $\mathbb{E}_{s\sim d^\pi} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a) \leq \widehat{A}_i + 2\epsilon\sqrt{|\mathcal{A}|}$.

Proof: Note that by Jensen's inequality, we have $\mathbb{E}_{s\sim d^{\pi,\nu}} |\widehat{Q}(s,a) - Q^\pi(s,a)| \leq \left\|\widehat{Q} - Q^\pi\right\|_{d^{\pi,\nu}}$, where we have dropped the subscripts involving i as we focus on a specific iteration. Denote $\pi^+(s) = \mathrm{argmax}_{a\in\mathcal{A}} Q^\pi(s,a)$. Now we have

$$\begin{aligned}\mathbb{E}_{s,a\sim d^\pi} A^\pi(s, \pi^+(s)) &= \mathbb{E}_{s\sim d^\pi}\Big[Q^\pi(s,\pi^+(s)) - \widehat{Q}(s,\pi^+(s)) + \widehat{Q}(s,\pi^+(s)) - \widehat{Q}(s,\widetilde{\pi}(s)) + \widehat{Q}(s,\widetilde{\pi}(s)) \\ &\qquad - \widehat{Q}(s,\pi(s)) + \widehat{Q}(s,\pi(s)) - Q^\pi(s,\pi(s))\Big] \\ &\leq \left\|\widehat{Q} - Q^\pi\right\|_{d^{\pi,\pi^+}} + \left\|\widehat{Q} - Q^\pi\right\|_{d^{\pi,\pi}} + \mathbb{E}_{s\sim d^\pi}[\widehat{Q}(s,\widetilde{\pi}(s)) - \widehat{Q}(s,\pi(s))],\end{aligned}$$

where we have used the inequality $\widehat{Q}(s,\pi^+(s)) \leq \widehat{Q}(s,\widetilde{\pi}(s))$ by the definition of $\widetilde{\pi}$. ■

Putting these two results together, we obtain the following useful corollary which gives a lower bound on the per-step improvement of the CPI algorithm for our choice of α_i.

Corollary 5.10. *At each iteration of Algorithm 2, we have* $V^{\pi_{i+1}} - V^{\pi_i} \geq (\widehat{A}_i - \epsilon)^2/8$.

Proof: From Theorem 5.8, we have

$$\begin{aligned}V^{\pi_{i+1}} - V^{\pi_i} &\geq \frac{\alpha}{1-\gamma}\mathbb{E}_{s\sim d^\pi} A^{\pi_i}(s, \widetilde{\pi}_i(s)) - \frac{2\alpha^2}{(1-\gamma)(1-\gamma(1-\alpha))} \\ &\geq \frac{\alpha}{1-\gamma}(\widehat{A}_i - \epsilon) - \frac{2\alpha^2}{(1-\gamma)(1-\gamma(1-\alpha))} && \text{(Lemma 5.9 and Line 2 in Algorithm 2)} \\ &\geq \frac{\alpha}{1-\gamma}\left(\widehat{A}_i - \epsilon - \frac{2\alpha}{1-\gamma}\right) \\ &\geq \frac{(\widehat{A}_i - \epsilon)^2}{8} && \text{(Using } \alpha_i = (\widehat{A}_i - \epsilon)(1-\gamma)/4\text{)}\end{aligned}$$

■

Putting these results together, we obtain the following overall convergence guarantee for the CPI algorithm.

Theorem 5.11 (Local optimality of CPI)**.** *Algorithm 2 terminates in at most* $8/\epsilon^2$ *steps and outputs a policy* π *satisfying* $\mathbb{E}_{s\sim d^\pi} \max_a A^\pi(s,a) \leq 3\epsilon$.

Proof: We can now prove the theorem. Note that each at each iteration, CPI either terminates or ensures that $\widehat{A}_i \geq 2\epsilon$. By Corollary 5.10, this implies that if the algorithm has performed k iterations, we must have

$$V^{\pi_k} - V^{\pi_0} \geq \frac{k\epsilon^2}{8}.$$

Since all policies take values in $[0,1]$, this means that the algorithm must terminate in at most $8\epsilon^2$ iterations. Furthermore, at termination, we have $\widehat{A}_k \leq 2\epsilon$. Combined with Lemma 5.9 and the accuracy of $\widehat{Q}_i$ in Line 2 in Algorithm 2 completes the proof. ■

Theorem 5.11 can be viewed as a local optimality guarantee in a sense. It shows that when CPI terminates, no improvements are possible to the returned policy π by updating it in the direction of one-step advantages. However, this does not necessarily imply that the value of π is close to $V^\star$. Indeed if the MDP has a chain like structure which we saw in the policy gradient lecture, and π advances towards the goal state with a small probability only, then it is easy to convince ourselves that no local improvements might be possible even when the policy is significantly suboptimal. However, similar to the policy gradient analysis, we can turn this local guarantee into a global one when the resulting policy's state distribution is sufficiently exploratory. We formalize this intuition next.

Theorem 5.12 (Global optimality of CPI). *Given any policy π satisfying $\mathbb{E}_{s\sim d^\pi} \max_{a\in\mathcal{A}} A^\pi(s,a) \leq \epsilon$, we have*

$$V^* - V^\pi \leq \frac{\epsilon}{1-\gamma}\left\|\frac{d^{\pi^*}}{d^\pi}\right\|_\infty \leq \frac{\epsilon}{(1-\gamma)^2}\left\|\frac{d^{\pi^*}}{d_0}\right\|_\infty,$$

where the ratio d^{π^*}/d^π represents elementwise division.

Proof: The proof is essentially contained in that of Theorem 4.11. By the performance difference lemma,

$$\begin{aligned} V^* - V^\pi &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi^*}} A^\pi(s,\pi^*(s)) \\ &\leq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi^*}} \max_{a\in\mathcal{A}} A^\pi(s,a) \\ &\leq \frac{1}{1-\gamma}\left\|\frac{d^{\pi^*}}{d^\pi}\right\|_\infty \mathbb{E}_{s\sim d^\pi} \max_{a\in\mathcal{A}} A^\pi(s,a). \end{aligned}$$

■

Consequently, if the initial distribution, or the visitation distribution of the final policy output by CPI is sufficiently exploratory relative to π^*, then we converge to a globally optimal policy.

It is informative to contrast CPI and policy gradient algorithms due to the similarity of their guarantees. Both provide local optimality guarantees. For CPI, the local optimality always holds, while for policy gradients it requires a smooth value function as a function of the policy parameters. If the distribution mismatch between an optimal policy and the output of the algorithm is not too large, then both algorithms further yield a near optimal policy. The similarities are not so surprising. Both algorithms operate by making local improvements to the current policy at each iteration, by inspecting its advantage function. The changes made to the policy are controlled using a stepsize parameter in both the approaches. It is the actual mechanism of the improvement which differs in the two cases. Policy gradients assume that the policy's reward is a differentiable function of the parameters, and hence make local improvements through gradient ascent. The differentiability is certainly an assumption and does not necessarily hold for all policy classes. An easy example is when the policy itself is not an easily differentiable function of its parameters. For instance, if the policy is parametrized by regression trees, then performing gradient updates can be challenging.

In CPI, on the other hand, the basic computational primitive required on the policy class is the ability to maximize the advantage function relative to the current policy. Notice that Algorithm 2 does not necessarily restrict to a policy class, such as a set of parametrized policies as in policy gradients. But in practice, we typically seek to find a policy from a restricted class which yields the largest improvement on the current policy. In this case, we can no longer set $\widetilde{\pi}_i(s) = \max_{a\in\mathcal{A}} \widehat{Q}_i(s,a)$, but instead seek $\widetilde{\pi}_i(s) = \operatorname{argmax}_{\pi\in\Pi} \mathbb{E}_{s\sim d^{\pi_i}} \widehat{Q}_i(s,\pi(s))$. The latter can be viewed as a weighted classification problem where the policies π are seen as probabilistic classifiers mapping states to actions, and the cost of prediction an action a on a state s is $-\widehat{Q}_i(s,a)$. If we collect a dataset of (s,a) samples with $s \sim d^{\pi_i}$ and uniformly random actions, then it is easy to show that the best improvement policy can be obtained by easy modifications of existing classification or regression algorithms. This property makes CPI extremely attractive. Any policy class over which efficient supervised learning algorithms exist can be adapted to reinforcement learning with performance guarantees.

A second important difference between CPI and policy gradients is in the notion of locality. Policy gradient updates are local in the parameter space, and we hope that this makes small enough changes to the state distribution that the new policy is indeed an improvement on the older one (for instance, when we invoke the performance difference lemma between successive iterates). While this is always true in expectation for correctly chosen stepsizes based on properties of stochastic gradient ascent on smooth functions, the variance of the algorithm and lack of robustness to suboptimal stepsizes can make the algorithm somewhat finicky. Indeed, there are a host of techniques in the literature to both lower the variance (through control variates) and explicitly control the state distribution mismatch between successive iterates of policy gradients (through trust region techniques). On the other hand, CPI explicitly controls the amount of perturbation to the state distribution by carefully mixing policies in a manner which does not drastically alter the trajectories with high probability. Indeed, this insight is central to the proof of CPI, and has been instrumental in several follow-ups, both in the direct policy improvement as well as policy gradient literature.

Bibliography

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement lear ning with a generative model. *Machine learning*, 91(3):325–349, 2013.

J. Andrew Bagnell and Jeff Schneider. Covariant policy search. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pages 1019–1024, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc. URL `http://dl.acm.org/citation.cfm?id=1630659.1630805`.

Richard Bellman. Dynamic programming and Lagrange multipliers. *Proceedings of the National Academy of Sciences*, 42(10):767–769, 1956.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandit algorithms with supervised learning guarantees. In *AISTATS*, 2011.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *The Journal of Machine Learning Research*, 3:213–231, 2003.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8:231–357, 2015. ISSN 1935-8237.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Found. Trends Mach. Learn.*, 10: 142–336, 2017. ISSN 1935-8237.

S. Kakade. A natural policy gradient. In *NIPS*, 2001.

Sham Kakade and John Langford. Approximately Optimal Approximate Reinforcement Learning. In *Proceedings of the 19th International Conference on Machine Learning*, volume 2, pages 267–274, 2002.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University of College London, 2003.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 2002.

Martin Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 1994.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F. Yang, and Yinyu Ye. Near-optimal time and sample complexities for for solving discounted markov decision process with a generative model. In *Advances in Neural Information Processing Systems 31*, 2018.

Satinder Singh and Richard Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 99, pages 1057–1063, 1999.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Appendix A

Concentration

Lemma A.1. *(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 μ. 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 A.2. *(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 $Var(X_i)$ denote the variance of X_i. If $X_i - EX_i \leq b$ for all i, then*

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

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

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

Lemma A.3 (Version of Freedman's inequality from Beygelzimer et al. [2011])**.** *Let $X_1, X_2, \ldots, X_T$ be a sequence of real-valued random variables adapted to the filtration $\mathcal{F}_i$. That is, X_i is measurable with respect to $\mathcal{F}_i$ and further assume that $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}]$. Define $S = \sum_{t=1}^T X_t$, $V = \sum_{t=1}^T \mathbb{E}[X_t^2 \mid \mathcal{F}_{t-1}]$ and let $X_t \leq R$ almost surely for all t. Then for any $\delta \in (0,1)$ and $\lambda \in [0, 1/R]$, with probability at least $1 - \delta$,*

$$S \leq (e-2)\lambda V + \frac{\ln(1/\delta)}{\lambda}.$$

In particular, choosing $\lambda = \min\left\{\frac{1}{R}, \sqrt{\ln(1/\delta)V}\right\}$, we get the Bernstein-style bound

$$S \leq 2\sqrt{V\ln\frac{1}{\delta}} + R\ln\frac{1}{\delta}.$$