# Chapter 3

# Optimism and Bellman Equations

In the previous chapter, we ran the policy gradient algorithm on two environments: CartPole and Pong. We observed that finding the optimal policy was much easier in CartPole than in Pong. This naturally raises the question: what makes some RL problems easier than others?

By "easy," we mean problems for which we can design algorithms that are efficient in both sample complexity (the amount of interaction required) and computational complexity (the amount of computation, e.g., GPU hours, needed).

This chapter will focus on: when can we design sample efficient RL algorithms? This chapter introduces two core algorithmic ideas in reinforcement learning theory: *optimism* and *bellman equations*. These principles underlie many sample-efficient algorithms across various settings, including Linear Bellman-Complete MDPs [5], Linear MDPs [6], Deterministic linear Q* [7], and more. The first attempt to generalize these ideas was made in [8]. A broader overview can be found in [9].

In this chapter, we focus on the simplest such setting—the *linear setting*—where the optimal value functions $V^*$ and $Q^*$ lie in a known linear function class.

**Definition 3.1** (Linear $V^*$ and $Q^*$)**.** *We assume that the optimal value functions admit a known linear feature representation:*

- ***Linear*** $Q^*$***:*** *There exists an unknown parameter vector $w^* \in \mathbb{R}^d$ and a known feature map $\phi : S \times A \to \mathbb{R}^d$ such that*

$$Q^*(s,a) = \langle w^*, \phi(s,a) \rangle \quad \textit{for all } (s,a) \in S \times A.$$

- ***Linear*** $V^*$***:*** *There exists an unknown parameter vector $\theta^* \in \mathbb{R}^d$ and a known feature map $\phi : S \to \mathbb{R}^d$ such that*

$$V^*(s) = \langle \theta^*, \phi(s) \rangle \quad \textit{for all } s \in S.$$

*We use the same notation $\phi$ for both feature maps, with the meaning clear from context.*

We study the problem of learning a near-optimal policy in the finite-horizon linear setting, where the agent interacts with the environment and seeks a *sample-efficient* algorithm: one that learns a good policy using as few episodes as possible.

**Definition 3.2** (Goal: Sample-Efficient Algorithm for Linear Setting)**.** *Let $\varepsilon, \delta \in (0,1)$ be accuracy and confidence parameters. The goal is to design an algorithm which, with probability at least $1-\delta$, outputs a policy $\hat{\pi}$ such that*

$$V^{\hat{\pi}}(s_0) \geq V^*(s_0) - \varepsilon,$$

*where $V^*$ denotes the value of an optimal policy and $s_0$ is the starting state. The algorithm is* sample efficient *if the number of environment interactions required is polynomial in the feature dimension $d$, number of actions $|A|$, horizon $H$, and parameters $\varepsilon^{-1}, \log \delta^{-1}$.*

## 3.1 Bellman Equation

A key observation behind our algorithm is the Bellman equation, which holds for all state-action pairs $(s_h, a_h)$:

$$Q_h^*(s_h, a_h) = \mathop{\mathbb{E}}_{\substack{s_{h+1} \sim T(s_h, a_h) \\ r_h \sim R(s_h, a_h)}} \left[ r_h + V_{h+1}^*(s_{h+1}) \right]. \tag{3.1}$$

This immediately implies:

$$\mathop{\mathbb{E}}_{\substack{s_{h+1} \sim T(s_h, a_h) \\ r_h \sim R(s_h, a_h)}} \left[ Q_h^*(s_h, a_h) - r_h - V_{h+1}^*(s_{h+1}) \right] = 0. \tag{3.2}$$

We are interested in analyzing this identity under the distribution induced by a policy, which we define next.

**Definition 3.3** (Trajectory Distribution $d^\pi$)**.** *Let $\pi$ be a policy. The distribution $d_h^\pi$ denotes the marginal over $(s_h, a_h, r_h, s_{h+1})$ at time $h$ under policy $\pi$, with the following sampling process:*

- *$s_0$ is the known starting state,*
- *$a_t \sim \pi(s_t)$, $s_{t+1} \sim T(s_t, a_t)$, and $r_t \sim R(s_t, a_t)$.*

*We write $d^\pi = \{d_0^\pi, \ldots, d_{H-1}^\pi\}$ to denote the full set of marginals over the trajectory.*

Taking expectations in Equation (3.2) over any distribution on $(s_h, a_h)$, including the trajectory distribution $d^\pi$ induced by a policy $\pi$, preserves the identity:

$$\mathop{\mathbb{E}}_{(s_h, a_h, r_h, s_{h+1}) \sim d_h^\pi} \left[ Q^*(s_h, a_h) - r_h - V_{h+1}^*(s_{h+1}) \right] = 0. \tag{3.3}$$

For brevity, henceforth, we will use $\mathbb{E}_{d_h^\pi}$ to denote the above expectation. Finally, note that given a policy $\pi$, it is straightforward to sample from $d^\pi$ by interacting with the environment: simply run $\pi$ for multiple episodes and collect the tuples $(s_h, a_h, r_h, s_{h+1})$ at each time step $h$.

## 3.2 Algorithm

We now present a sample-efficient algorithm that iteratively constructs policies that are (a) *optimistic*, and (b) *approximately satisfy Bellman equation on past data*. At each iteration $t$, the algorithm uses data collected under previous policies $\pi_1, \ldots, \pi_t$ to construct a new policy $\pi_{t+1}$ that maximizes estimated value while ensuring that its Bellman residual (Equation (3.3)) is small under the empirical trajectory distributions from earlier policies.

Let $\hat{d}_h^{\pi_k}$ denote the empirical distribution over $n_{\text{emp}} = \text{poly}(d, H)$ transitions $(s_h, a_h, r_h, s_{h+1})$ collected from executing policy $\pi_k$. The number of samples collected per policy $n_{\text{emp}}$ and the constraint parameter $\varepsilon_{\text{cons}}$ will be chosen later in the analysis.

**Algorithm 2:** Optimistic Algorithm

**for** $t = 1, 2, \ldots, n_{rounds}$ **do**

Solve the following optimization problem:

$$\max_{\theta, w} \langle \theta, \phi(s_0) \rangle$$

$$\text{s.t.} \quad \sum_{k=1}^{t-1} \mathop{\mathbb{E}}_{\hat{d}_h^{\pi_k}} \left[ \langle w, \phi(s_h, a_h) \rangle - r_h - \langle \theta, \phi(s_{h+1}) \rangle \right]^2 \leq \frac{\varepsilon^2}{\text{poly}(d, H)}$$

Here, $\|\theta\|_2$, $\|w\|_2 \leq 1$ are norm constrained and satisfy $\langle \theta, \phi(s) \rangle = \max_a \langle w, \phi(s, a) \rangle$ for all states $s$.

Let $\theta_t, w_t$ be the resulting parameters. Define $\pi_t$ as the resulting policy defined by $\pi_t(s) = \text{argmax}_a \langle w_t, \phi(s, a) \rangle$.

Execute $\pi_t$ for $n_{\text{emp}}$ episodes, and use the collected data to construct empirical distributions $\hat{d}^{\pi_t} = \{\hat{d}_h^{\pi_t}\}_{h=0}^{H-1}$.;

**return** the best policy $\pi_t$ observed so far.;

## 3.3 Optimization Constraint in Linear Form

We begin by rewriting the constraint in our optimization program using the linear representations of the value functions. Recall that we denote the time-indexed value functions as $Q_h^\pi$ and $V_h^\pi$, where $h \in \{0, \ldots, H-1\}$. However, to keep notation light, we will suppress the dependence on $h$ in the remainder of

this proof, with the understanding that the analysis applies separately at each time step.

At round $t$, let $\theta_t, w_t$ be the resulting parameters, and define $\pi_t$ as the resulting policy defined by $\pi_t(s) = \text{argmax}_a \langle w_t, \phi(s,a)\rangle$. Define the concatenated parameter and feature vectors:

$$W(\pi_t) := [w_t, \theta_t] \in \mathbb{R}^{2d},$$
$$X(\pi_t) := \mathop{\mathbb{E}}_{d_h^{\pi_t}} [\phi(s_h, a_h), -\phi(s_{h+1})] \in \mathbb{R}^{2d}.$$

Here, recall the notation $\mathbb{E}_{d_h^{\pi_t}}$ is shorthand for the expectation over samples $(s_h, a_h, s_{h+1}) \sim d_h^{\pi_t}$. The expected Bellman optimality equation for a candidate policy $\pi_t$ at time step $h$ is:

$$\mathop{\mathbb{E}}_{d^{\pi_t}} [\langle w_t, \phi(s_h,a_h)\rangle - r_h - \langle \theta_t, \phi(s_{h+1})\rangle] \tag{3.4}$$
$$= \mathop{\mathbb{E}}_{d^{\pi_t}} [\langle w_t, \phi(s_h,a_h)\rangle - \langle \theta_t, \phi(s_{h+1})\rangle - Q^*(s_h,a_h) + V^*(s_{h+1})]$$
$$= \langle W(\pi_t) - W(\pi^*), X(\pi_t)\rangle, \tag{3.5}$$

where the first equality uses the Bellman identity for the optimal policy (Equation (3.3)) to replace $r_h$ with $Q^*(s_h,a_h) - V^*(s_{h+1})$ in expectation.

## 3.4 Optimism: Bounding Regret

To analyze the regret, we relate the suboptimality of the current policy $\pi_t$ to the difference between its parameters and those of the optimal value functions. The key idea is optimism: the algorithm selects $\theta_t, w_t$ to maximize $\langle \theta_t, \phi(s_0)\rangle$, which serves as an optimistic estimate of $V^{\pi_t}(s_0)$. Although this estimate may be inaccurate, the algorithm proceeds by acting according to $\pi_t$.

**Lemma 3.4** (Regret Decomposition via Optimism)**.** *Let the policy $\pi_t$, and vectors $W(\pi_t), X(\pi_t) \in \mathbb{R}^{2d}$ be as defined above. Then:*

$$V^*(s_0) - V^{\pi_t}(s_0) \le \sum_{h=0}^{H-1} \left| \langle W(\pi_t) - W(\pi^*), X_h(\pi_t)\rangle \right|.$$

*Proof.* Throughout, we use that $a_h = \text{argmax}_a \langle w_t, \phi(s_h, a)\rangle$. The claim follows

from:

$$
\begin{aligned}
& V^*(s_0) - V^{\pi_t}(s_0) \\
& \leq \langle \theta_t, \phi(s_0) \rangle - V^{\pi_t}(s_0) && \text{(follows from optimism)} \\
& = \langle w_t, \phi(s_0, a_0) \rangle - \mathop{\mathbb{E}}_{d^{\pi_t}} \left[ \sum_{h=0}^{H-1} r_h \right] && \text{(since } \langle \theta_t, \phi(s_0) \rangle = \langle w_t, \phi(s_0, a_0) \rangle \text{)} \\
& = \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{d_h^{\pi_t}} \left[ \langle w_t, \phi(s_h, a_h) \rangle - r_h - \langle w_t, \phi(s_{h+1}, a_{h+1}) \rangle \right] && \text{(by telescoping sum)} \\
& = \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{d_h^{\pi_t}} \left[ \langle w_t, \phi(s_h, a_h) \rangle - r_h - \langle \theta_t, \phi(s_{h+1}) \rangle \right] \\
& && \text{(since } \langle \theta_t, \phi(s_{h+1}) \rangle = \langle w_t, \phi(s_h, a_h) \rangle \text{)} \\
& \leq \sum_{h=0}^{H-1} \left| \langle W(\pi_t) - W(\pi^*), X_h(\pi_t) \rangle \right|. && \text{(follows from Equation (3.5))}
\end{aligned}
$$

□

## 3.5 Exploration: Bounding the Number of Rounds

The final step is to bound how many rounds $n_{\text{rounds}}$ are needed before the algorithm identifies a policy $\pi_t$ with small suboptimality. We provide a simplified argument, which can be improved to obtain tighter bounds.

The key idea is that for large enough $n_{\text{emp}}$, the empirical distribution $\hat{d}_h^{\pi_k}$ concentrates around the true distribution $d_h^{\pi_k}$. Then, from the regret decomposition lemma above (Lemma 3.4), we know that if the residual term $\langle W(\pi_k) - W(\pi^*), X_h(\pi_k) \rangle$ is zero for all $h$, then the policy $\pi_k$ is optimal. Otherwise, a nonzero residual indicates that $\pi_k$ adds a new constraint that helps eliminate a portion of the hypothesis space.

More precisely, these constraints are linear in the difference vector $W(\pi) - W(\pi^*)$, and each non-redundant policy introduces a constraint that is at least $\varepsilon$-independent (Definition 3.5) of the previous ones. Since the vector space has finite dimension $2d$, only a bounded number of such $\varepsilon$-independent constraints can exist. This implies that the algorithm can only generate a finite number of meaningfully distinct (and suboptimal) policies before it must identify the optimal one.

Thus, it suffices to analyze the following situation: how many times can it happen that

$$
\begin{aligned}
\sum_{k=1}^{t-1} \left( \langle W(\pi_t) - W(\pi^*), X_h(\pi_k) \rangle \right)^2 &\leq \varepsilon^2 \\
\left( \langle W(\pi_t) - W(\pi^*), X_h(\pi_t) \rangle \right)^2 &> \varepsilon^2
\end{aligned}
$$

This question—how many times an $\varepsilon$-independent constraint can arise—was first analyzed by Russo and Van Roy [10].

**Definition 3.5** ($\varepsilon$-independence)**.** *Let $S_d = \{\theta \in R^d : \|\theta\|_2 \leq 1\}$. We say $x_n \in S_d$ is $\varepsilon$-independent of sequence $\{x_1, x_2, \ldots, x_{n-1}\}$ if there exists $\theta, \theta^* \in S_d$ such that*

$$\sum_{i=1}^{n-1} \left(\langle \theta, x_i \rangle - \langle \theta^*, x_i \rangle\right)^2 \leq \varepsilon^2$$
$$\left(\langle \theta, x_n \rangle - \langle \theta^*, x_n \rangle\right)^2 > \varepsilon^2$$

We now bound the number of such $\varepsilon$-independent vectors that can appear in a sequence.

**Lemma 3.6** (Bound on Number of $\varepsilon$-Independent Constraints)**.** *Let $\{x_1, \ldots, x_n\} \subseteq S_d$ be a sequence such that each $x_k$ is $\varepsilon$-independent of the previous vectors in the sense of Definition 3.5. Then, the number $n$ of such vectors is at most*

$$n = O\left(d \cdot \log\left(\frac{1}{\varepsilon}\right)\right).$$

*Proof.* More generally, assume that $\|x_k\|_2 \leq B_\phi$ for all $k$, and that $\|\theta - \theta^*\|_2 \leq 2B_\theta$. In our case, both $B_\phi$ and $B_\theta$ are 1.

Let $V_n := \sum_{i=1}^{n-1} x_i x_i^\top + \lambda I$ for some regularization $\lambda = \varepsilon^2/(2B_\theta)^2$. If $x_n$ is $\varepsilon$-independent of the previous $x_1, \ldots, x_{n-1}$, then there exists $\theta, \theta^*$ with $\|\theta - \theta^*\|_2 \leq 2B_\theta$ such that:

$$\sum_{i=1}^{n-1} \left(\langle \theta - \theta^*, x_i \rangle\right)^2 \leq \varepsilon^2, \quad \text{but} \quad \left(\langle \theta - \theta^*, x_n \rangle\right)^2 > \varepsilon^2.$$

This implies:

$$\begin{aligned}
\varepsilon &\leq \left\{ (\theta - \theta^*)^\top x_n : (\theta - \theta^*)^\top \left(\sum_{i=1}^{n-1} x_i x_i^\top\right)(\theta - \theta^*) \leq \varepsilon^2 \quad \text{and} \quad (\theta - \theta^*)^\top I\,(\theta - \theta^*) \leq (2B_\theta)^2 \right\} \\
&\leq \max_\rho \left\{ \rho^\top x_n : \rho^\top \left(\sum_{i=1}^{n-1} x_i x_i^\top\right)\rho \leq \varepsilon^2 \quad \text{and} \quad \rho^\top I \rho \leq (2B_\theta)^2 \right\} \\
&\leq \max_\rho \left\{ \rho^\top x_n : \rho^\top \left(\sum_{i=1}^{n-1} x_i x_i^\top + \lambda I\right)\rho \leq 2\varepsilon^2 \right\} \\
&= \sqrt{2\varepsilon^2}\|x_n\|_{V_n^{-1}}
\end{aligned}$$

And therefore,

$$\|x_n\|^2_{V_n^{-1}} = x_n^\top V_n^{-1} x_n \geq \frac{1}{2}.$$

We now analyze the growth of $\det(V_n)$ using the matrix determinant lemma:

$$\begin{aligned}\det(V_n) &= \det(V_{n-1})\left(1 + x_{n-1}^\top V_{n-1}^{-1} x_{n-1}\right) \\ &\geq \det(V_{n-1}) \cdot \frac{3}{2}.\end{aligned}$$

Iterating gives:

$$\det(V_n) \geq \det(\lambda I) \cdot \left(\frac{3}{2}\right)^{n-1} = \lambda^d \left(\frac{3}{2}\right)^{n-1}.$$

On the other hand, since $\operatorname{trace}(V_n) \leq nB_\phi^2 + d\lambda$, we apply the AM–GM inequality:

$$\det(V_n) \leq \left(\frac{\operatorname{trace}(V_n)}{d}\right)^d \leq \left(\frac{nB_\phi^2}{d} + \lambda\right)^d.$$

Equating the lower and upper bounds on $\det(V_n)$ gives:

$$\lambda^d \left(\frac{3}{2}\right)^{n-1} \leq \left(\frac{nB_\phi^2}{d} + \lambda\right)^d.$$

Solving this inequality yields the desired bound:

$$n = O\left(d \cdot \log\left(\frac{B_\theta^2 B_\phi^2}{\varepsilon}\right)\right).$$

□