Chapter 8

Linearly Parameterized MDPs

In this chapter, we consider learning and exploration in linearly parameterized MDPs—the linear MDP. Linear MDP generalizes tabular MDPs into MDPs with potentially infinitely many state and action pairs.

This chapter largely follows the model and analysis first provided in [Jin et al., 2020].

8.1 Setting

We consider episodic finite horizon MDP with horizon H, $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \{r_h\}_h, \{P_h\}_h, H, s_0\}$, where s_0 is a fixed initial state, $r_h : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ and $P_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ are time-dependent reward function and transition kernel. Note that for time-dependent finite horizon MDP, the optimal policy will be time-dependent as well. For simplicity, we overload notations a bit and denote $\pi = \{\pi_0, \ldots, \pi_{H-1}\}$, where each $\pi_h : \mathcal{S} \mapsto \mathcal{A}$. We also denote $V^\pi := V_0^\pi(s_0)$, i.e., the expected total reward of π starting at $h = 0$ and s_0.

We define the learning protocol below. Learning happens in an episodic setting. Every episode k, learner first proposes a policy π^k based on all the history information up to the end of episode $k-1$. The learner then executes π^k in the underlying MDP to generate a single trajectory $\tau^k = \{s_h^k, a_h^k\}_{h=0}^{H-1}$ with $a_h = \pi_h^k(s_h^k)$ and $s_{h+1}^k \sim P_h(\cdot|s_h^k, a_h^k)$. The goal of the learner is to minimize the following cumulative regret over N episodes:

$$\text{Regret} := \mathbb{E}\left[\sum_{k=0}^{K-1}\left(V^\star - V^{\pi^k}\right)\right],$$

where the expectation is with respect to the randomness of the MDP environment and potentially the randomness of the learner (i.e., the learner might make decisions in a randomized fashion).

8.1.1 Low-Rank MDPs and Linear MDPs

Note that here we do not assume $\mathcal{S}$ and $\mathcal{A}$ are finite anymore. Indeed in this note, both of them could be continuous. Without any further structural assumption, the lower bounds we saw in the Generalization Lecture forbid us to get a polynomially regret bound.

The structural assumption we make in this note is a linear structure in both reward and the transition.

Definition 8.1 (Linear MDPs)**.** Consider transition $\{P_h\}$ and $\{r_h\}_h$. A linear MDP has the following structures on r_h

and P_h:

$$r_h(s,a) = \theta_h^\star \cdot \phi(s,a), \quad P_h(\cdot|s,a) = \mu_h^\star \phi(s,a), \forall h$$

where ϕ is a known state-action feature map $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, and $\mu_h^\star \in \mathbb{R}^{|\mathcal{S}| \times d}$. Here $\phi, \theta_h^\star$ are **known** to the learner, while $\mu^\star$ is unknown. We further assume the following norm bound on the parameters: (1) $\sup_{s,a} \|\phi(s,a)\|_2 \leq 1$, (2) $\|v^\top \mu_h^\star\|_2 \leq \sqrt{d}$ for any v such that $\|v\|_\infty \leq 1$, and all h, and (3) $\|\theta_h^\star\|_2 \leq W$ for all h. We assume $r_h(s,a) \in [0,1]$ for all h and s, a.

The model essentially says that the transition matrix $P_h \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}||\mathcal{A}|}$ has rank at most d, and $P_h = \mu_h^\star \Phi$. where $\Phi \in \mathbb{R}^{d \times |\mathcal{S}||\mathcal{A}|}$ and each column of Φ corresponds to $\phi(s,a)$ for a pair $s, a \in \mathcal{S} \times \mathcal{A}$.

Linear Algebra Notations For real-valued matrix A, we denote $\|A\|_2 = \sup_{x:\|x\|_2=1} \|Ax\|_2$ which denotes the maximum singular value of A. We denote $\|A\|_F$ as the Frobenius norm $\|A\|_F^2 = \sum_{i,j} A_{i,j}^2$ where $A_{i,j}$ denotes the i,j'th entry of A. For any Positive Definite matrix Λ, we denote $x^\top \Lambda x = \|x\|_\Lambda^2$. We denote $\det(A)$ as the determinant of the matrix A. For a PD matrix Λ, we note that $\det(\Lambda) = \prod_{i=1}^d \sigma_i$ where σ_i is the eigenvalues of Λ. For notation simplicity, during inequality derivation, we will use $\lesssim, \eqsim$ to suppress all absolute constants. We will use $\widetilde{O}$ to suppress all absolute constants and log terms.

8.2 Planning in Linear MDPs

We first study how to do value iteration in linear MDP if μ is given.

We start from $Q_{H-1}^\star(s,a) = \theta_{H-1}^\star \cdot \phi(s,a)$, and $\pi_{H-1}^\star(s) = \operatorname{argmax}_a Q_{H-1}^\star(s,a) = \operatorname{argmax}_a \theta_{H-1}^\star \cdot \phi(s,a)$, and $V_{H-1}^\star(s) = \operatorname{argmax}_a Q_{H-1}^\star(s,a)$.

Now we do dynamic programming from $h+1$ to h:

$$Q_h^\star(s,a) = \theta_h^\star \cdot \phi(s,a) + \mathbb{E}_{s \sim P_h(\cdot|s,a)} V_{h+1}^\star(s') = \theta_h^\star \cdot \phi(s,a) + P_h(\cdot|s,a) \cdot V_{h+1}^\star = \theta_h^\star \cdot \phi(s,a) + (\mu_h^\star \phi(s,a))^\top V_{h+1}^\star \tag{0.1}$$

$$= \phi(s,a) \cdot \left(\theta_h^\star + (\mu_h^\star)^\top V_{h+1}^\star\right) = \phi(s,a) \cdot w_h, \tag{0.2}$$

where we denote $w_h := \theta_h^\star + (\mu_h^\star)^\top V_{h+1}^\star$. Namely we see that $Q_h^\star(s,a)$ is a linear function with respect to $\phi(s,a)$! We can continue by defining $\pi_h^\star(s) = \operatorname{argmax}_a Q_h^\star(s,a)$ and $V_h^\star(s) = \max_a Q_h^\star(s,a)$.

At the end, we get a sequence of linear $Q^\star$, i.e., $Q_h^\star(s,a) = w_h \cdot \phi(s,a)$, and the optimal policy is also simple, $\pi_h^\star(s) = \operatorname{argmax}_a w_h \cdot \phi(s,a)$, for all $h = 0, \ldots, H-1$.

One key property of linear MDP is that a Bellman Backup of any function $f : \mathcal{S} \mapsto \mathbb{R}$ is a linear function with respect to $\phi(s,a)$. We summarize the key property in the following claim.

Claim 8.2. *Consider any arbitrary function $f : \mathcal{S} \mapsto [0,H]$. At any time step $h \in [0, \ldots H-1]$, there must exist a $w \in \mathbb{R}^d$, such that, for all $s, a \in \mathcal{S} \times \mathcal{A}$:*

$$r_h(s,a) + P_h(\cdot|s,a) \cdot f = w^\top \phi(s,a).$$

The proof of the above claim is essentially the Eq. 0.1.

8.3 Learning Transition using Ridge Linear Regression

In this section, we consider the following simple question: given a dataset of state-action-next state tuples, how can we learn the transition P_h for all h?

Note that $\mu^\star \in \mathbb{R}^{|\mathcal{S}| \times d}$. Hence explicitly writing down and storing the parameterization $\mu^\star$ takes time at least $|\mathcal{S}|$. We show that we can represent the model in a non-parametric way.

We consider a particular episode n. Similar to Tabular-UCBVI, we learn a model at the very beginning of the episode n using all data from the previous episodes (episode 1 to the end of the episode $n-1$). We denote such dataset as:

$$\mathcal{D}_h^n = \left\{s_h^i, a_h^i, s_{h+1}^i\right\}_{i=0}^{n-1}.$$

We maintain the following statistics using $\mathcal{D}_h^n$:

$$\Lambda_h^n = \sum_{i=0}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I,$$

where $\lambda \in \mathbb{R}^+$ (it will be set to 1 eventually, but we keep it here for generality).

To get intuition of Λ^n, think about the tabular setting where $\phi(s,a)$ is a one-hot vector (zeros everywhere except that the entry corresponding to (s,a) is one). Then Λ_h^n is a diagonal matrix and the diagonal entry contains $N^n(s,a)$—the number of times (s,a) has been visited.

We consider the following multi-variate linear regression problem. Denote $\delta(s)$ as a one-hot vector that has zero everywhere except that the entry corresponding to s is one. Denote $\epsilon_h^i = P(\cdot|s_h^i, a_h^i) - \delta(s_{h+1}^i)$. Conditioned on history $\mathcal{H}_h^i$ (history $\mathcal{H}_h^i$ denotes all information from the very beginning of the learning process up to and including (s_h^i, a_h^i)), we have:

$$\mathbb{E}\left[\epsilon_h^i | \mathcal{H}_h^i\right] = 0,$$

simply because s_{h+1}^i is sampled from $P_h(\cdot|s_h^i, a_h^i)$ conditioned on (s_h^i, a_h^i). Also note that $\|\epsilon_h^i\|_1 \le 2$ for all h, i.

Since $\mu_h^\star \phi(s_h^i, a_h^i) = P_h(\cdot|s_h^i, a_h^i)$, and $\delta(s_{h+1}^i)$ is an unbiased estimate of $P_h(\cdot|s_h^i, a_h^i)$ conditioned on s_h^i, a_h^i, it is reasonable to learn $\mu^\star$ via regression from $\phi(s_h^i, a_h^i)$ to $\delta(s_{h+1}^i)$. This leads us to the following ridge linear regression:

$$\widehat{\mu}_h^n = \text{argmin}_{\mu \in \mathbb{R}^{|\mathcal{S}| \times d}} \sum_{i=0}^{n-1} \left\|\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\right\|_2^2 + \lambda \|\mu\|_F^2.$$

Ridge linear regression has the following closed-form solution:

$$\widehat{\mu}_h^n = \sum_{i=0}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1} \tag{0.3}$$

Note that $\widehat{\mu}_h^n \in \mathbb{R}^{|\mathcal{S}| \times d}$, so we never want to explicitly store it. Note that we will always use $\widehat{\mu}_h^n$ together with a specific s, a pair and a value function V (think about value iteration case), i.e., we care about $\widehat{P}_h^n(\cdot|s,a) \cdot V := (\widehat{\mu}_h^n \phi(s,a)) \cdot V$, which can be re-written as:

$$\widehat{P}_h^n(\cdot|s,a) \cdot V := (\widehat{\mu}_h^n \phi(s,a)) \cdot V = \phi(s,a)^\top \sum_{i=0}^{n-1} (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i) V(s_{h+1}^i),$$

where we use the fact that $\delta(s)^\top V = V(s)$. Thus the operator $\widehat{P}_h^n(\cdot|s,a)\cdot V$ simply requires storing all data and can be computed via simple linear algebra and the computation complexity is simply $\text{poly}(d,n)$—no poly dependency on $|\mathcal{S}|$.

Let us calculate the difference between $\widehat{\mu}_h^n$ and $\mu_h^\star$.

Lemma 8.3 (Difference between $\widehat{\mu}_h$ and $\mu_h^\star$)**.** *For all n and h, we must have:*

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda\mu_h^\star\left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1}\epsilon_h^i\phi(s_h^i,a_h^i)^\top\left(\Lambda_h^n\right)^{-1}.$$

Proof: We start from the closed-form solution of $\widehat{\mu}_h^n$:

$$\begin{aligned}
\widehat{\mu}_h^n &= \sum_{i=0}^{n-1}\delta(s_{h+1}^i)\phi(s_h^i,a_h^i)^\top(\Lambda_h^n)^{-1} = \sum_{i=0}^{n-1}(P(\cdot|s_h^i,a_h^i)+\epsilon_h^n)\phi(s_h^i,a_h^i)^\top(\Lambda_h^n)^{-1}\\
&= \sum_{i=0}^{n-1}(\mu_h^\star\phi(s_h^i,a_h^i)+\epsilon_h^i)\phi(s_h^i,a_h^i)^\top(\Lambda_h^n)^{-1} = \sum_{i=0}^{n-1}\mu_h^*\phi(s_h^i,a_h^i)\phi(s_h^i,a_h^i)^\top(\Lambda_h^n)^{-1} + \sum_{i=0}^{n-1}\epsilon_h^i\phi(s_h^i,a_h^i)^\top(\Lambda_h^n)^{-1}\\
&= \sum_{i=0}^{n-1}\mu_h^\star\phi(s_h^i,a_h^i)\phi(s_h^i,a_h^i)^\top(\Lambda_h^n)^{-1} + \sum_{i=0}^{n-1}\epsilon_h^i\phi(s_h^i,a_h^i)^\top(\Lambda_h^n)^{-1}\\
&= \mu_h^\star(\Lambda_h^n-\lambda I)(\Lambda_h^n)^{-1} + \sum_{i=0}^{n-1}\epsilon_h^i\phi(s_h^i,a_h^i)^\top(\Lambda_h^n)^{-1} = \mu_h^* - \lambda\mu_h^\star(\Lambda_h^n)^{-1} + \sum_{i=0}^{n-1}\epsilon_h^i\phi(s_h^i,a_h^i)^\top(\Lambda_h^n)^{-1}.
\end{aligned}$$

Rearrange terms, we conclude the proof. ■

Lemma 8.4. *Fix $V:\mathcal{S}\mapsto[0,H]$. For all n and $s,a\in\mathcal{S}\times\mathcal{A}$, and h, with probability at least $1-\delta$, we have:*

$$\left\|\sum_{i=0}^{n-1}\phi(s_h^i,a_h^i)(V^\top\epsilon_h^i)\right\|_{(\Lambda_h^n)^{-1}} \le 3H\sqrt{\ln\frac{H\det(\Lambda_h^n)^{1/2}\det(\lambda I)^{-1/2}}{\delta}}.$$

Proof: We first check the noise terms $\{V^\top\epsilon_h^i\}_{h,i}$. Since V is independent of the data (it's a pre-fixed function), and by linear property of expectation, we have:

$$\mathbb{E}\left[V^\top\epsilon_h^i|\mathcal{H}_h^i\right]=0,\quad |V^\top\epsilon_h^i|\le\|V\|_\infty\|\epsilon_h^i\|_1\le 2H,\forall h,i.$$

Hence, this is a Martingale difference sequence. Using the Self-Normalized vector-valued Martingale Bound (Lemma A.9), we have that for all n, with probability at least $1-\delta$:

$$\left\|\sum_{i=0}^{n-1}\phi(s_h^i,a_h^i)(V^\top\epsilon_h^i)\right\|_{(\Lambda_h^n)^{-1}} \le 3H\sqrt{\ln\frac{\det(\Lambda_h^n)^{1/2}\det(\lambda I)^{-1/2}}{\delta}}.$$

Apply union bound over all $h\in[H]$, we get that with probability at least $1-\delta$, for all n,h:

$$\left\|\sum_{i=0}^{n-1}\phi(s_h^i,a_h^i)(V^\top\epsilon_h^i)\right\|_{(\Lambda_h^n)^{-1}} \le 3H\sqrt{\ln\frac{H\det(\Lambda_h^n)^{1/2}\det(\lambda I)^{-1/2}}{\delta}}. \tag{0.4}$$

■

8.4 Uniform Convergence via Covering

Now we take a detour first and consider how to achieve a uniform convergence result over a function class $\mathcal{F}$ that contains infinitely many functions. Previously we know how to get uniform convergence if $\mathcal{F}$ is finite—we simply do a union bound. However, when $\mathcal{F}$ contains infinitely many functions, we cannot simply apply a union bound. We will use the covering argument here.

Consider the following ball with radius R: $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R \in \mathbb{R}^+\}$. Fix an ϵ. An ϵ-net $\mathcal{N}_\epsilon \subset \Theta$ is a set such that for any $\theta \in \Theta$, there exists a $\theta' \in \mathcal{N}_\epsilon$, such that $\|\theta - \theta'\|_2 \leq \epsilon$. We call the smallest ϵ-net as ϵ-cover. Abuse notations a bit, we simply denote $\mathcal{N}_\epsilon$ as the ϵ-cover.

The ϵ-covering number is the size of ϵ-cover $\mathcal{N}_\epsilon$. We define the covering dimension as $\ln(|\mathcal{N}_\epsilon|)$

Lemma 8.5. *The ϵ-covering number of the ball $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R \in \mathbb{R}^+\}$ is upper bounded by $(1 + 2R/\epsilon)^d$.*

We can extend the above definition to a function class. Specifically, we look at the following function. For a triple of (w, β, Λ) where $w \in \mathbb{R}^d$ and $\|w\|_2 \leq L$, $\beta \in [0, B]$, and Λ such that $\sigma_{\min}(\Lambda) \geq \lambda$, we define $f_{w,\beta,\Lambda} : \mathcal{S} \mapsto [0, R]$ as follows:

$$f_{w,\beta,\Lambda}(s) = \min\left\{\max_a \left(w^\top \phi(s,a) + \beta\sqrt{\phi(s,a)^\top \Lambda^{-1}\phi(s,a)}\right),\ H\right\}, \forall s \in \mathcal{S}. \tag{0.5}$$

We denote the function class $\mathcal{F}$ as:

$$\mathcal{F} = \{f_{w,\beta,\Lambda} : \|w\|_2 \leq L, \beta \in [0, B], \sigma_{\min}(\Lambda) \geq \lambda\}. \tag{0.6}$$

Note that $\mathcal{F}$ contains infinitely many functions as the parameters are continuous. However we will show that it has finite covering number that scales exponentially with respect to the number of parameters in (w, β, Λ).

Why do we look at $\mathcal{F}$? As we will see later in this chapter $\mathcal{F}$ contains all possible $\widehat{Q}_h$ functions one could encounter during the learning process.

Lemma 8.6 (ϵ-covering dimension of $\mathcal{F}$)**.** *Consider $\mathcal{F}$ defined in Eq. 0.6. Denote its ϵ-cover as $\mathcal{N}_\epsilon$ with the ℓ_∞ norm as the distance metric, i.e., $d(f_1, f_2) = \|f_1 - f_2\|_\infty$ for any $f_1, f_2 \in \mathcal{F}$. We have that:*

$$\ln(|\mathcal{N}_\epsilon|) \leq d\ln(1 + 6L/\epsilon) + \ln(1 + 6B/(\sqrt{\lambda}\epsilon)) + d^2\ln(1 + 18B^2\sqrt{d}/(\lambda\epsilon^2)).$$

Note that the ϵ-covering dimension scales quadratically with respect to d.

Proof: We start from building a net over the parameter space (w, β, Λ), and then we convert the net over parameter space to an ϵ-net over $\mathcal{F}$ under the ℓ_∞ distance metric.

We pick two functions that corresponding to parameters (w, β, Λ) and $(\hat{w}, \hat{\beta}, \widehat{\Lambda})$.

$$
\begin{aligned}
|f(s) - \hat{f}(s)| &\leq \left|\max_a \left(w^\top \phi(s,a) + \beta\sqrt{\phi(s,a)^\top \Lambda^{-1}\phi(s,a)}\right) - \max_a \left(\hat{w}^\top \phi(s,a) + \hat{\beta}\sqrt{\phi(s,a)^\top \hat{\Lambda}^{-1}\phi(s,a)}\right)\right| \\
&\leq \max_a \left|\left(w^\top \phi(s,a) + \beta\sqrt{\phi(s,a)^\top \Lambda^{-1}\phi(s,a)}\right) - \left(\hat{w}^\top \phi(s,a) + \hat{\beta}\sqrt{\phi(s,a)^\top \hat{\Lambda}^{-1}\phi(s,a)}\right)\right| \\
&\leq \max_a \left|(w-\hat{w})^\top \phi(s,a)\right| + \max_a \left|(\beta - \hat{\beta})\sqrt{\phi(s,a)^\top \Lambda^{-1}\phi(s,a)}\right| \\
&\qquad + \max_a \left|\hat{\beta}(\sqrt{\phi(s,a)^\top \Lambda^{-1}\phi(s,a)} - \sqrt{\phi(s,a)^\top \hat{\Lambda}^{-1}\phi(s,a)})\right| \\
&\leq \|w - \hat{w}\|_2 + |\beta - \hat{\beta}|/\sqrt{\lambda} + B\sqrt{\left|\phi(s,a)^\top (\Lambda^{-1} - \hat{\Lambda}^{-1})\phi(s,a)\right|} \\
&\leq \|w - \hat{w}\|_2 + |\beta - \hat{\beta}|/\sqrt{\lambda} + B\sqrt{\|\Lambda^{-1} - \hat{\Lambda}^{-1}\|_F}
\end{aligned}
$$

Note that Λ^{-1} is a PD matrix with $\sigma_{\max}(\Lambda^{-1}) \leq 1/\lambda$.

Now we consider the $\epsilon/3$-Net $\mathcal{N}_{\epsilon/3,w}$ over $\{w : \|w\|_2 \leq L\}$, $\sqrt{\lambda}\epsilon/3$-net $\mathcal{N}_{\sqrt{\lambda}\epsilon/3,\beta}$ over interval $[0, B]$ for β, and $\epsilon^2/(9B^2)$-net $\mathcal{N}_{\epsilon^2/(9B),\Lambda}$ over $\{\Lambda : \|\Lambda\|_F \leq \sqrt{d}/\lambda\}$. The product of these three nets provide a ϵ-cover for $\mathcal{F}$, which means that that size of the ϵ-net $\mathcal{N}_\epsilon$ for $\mathcal{F}$ is upper bounded as:

$$
\begin{aligned}
\ln|\mathcal{N}_\epsilon| &\leq \ln|\mathcal{N}_{\epsilon/3,w}| + \ln|\mathcal{N}_{\sqrt{\lambda}\epsilon/3,\beta}| + \ln|\mathcal{N}_{\epsilon^2/(9B^2),\Lambda}| \\
&\leq d\ln(1 + 6L/\epsilon) + \ln(1 + 6B/(\sqrt{\lambda}\epsilon)) + d^2\ln(1 + 18B^2\sqrt{d}/(\lambda\epsilon^2)).
\end{aligned}
$$

■

Remark Covering gives a way to represent the complexity of function class (or hypothesis class). Relating to VC, covering number is upper bound roughly by $\exp(d)$ with d being the VC-dimension. However, there are cases where VC-dimensional is infinite, but covering number if finite.

Now we can build a uniform convergence argument for all $f \in \mathcal{F}$.

Lemma 8.7 (Uniform Convergence Results)**.** *Set $\lambda = 1$. Fix $\delta \in (0,1)$. For all n, h, all s, a, and all $f \in \mathcal{F}$, with probability at least $1 - \delta$, we have:*

$$
\left|\left(\widehat{P}_h^n(\cdot|s,a) - P(\cdot|s,a)\right) \cdot f\right| \lesssim H\,\|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \left(\sqrt{d\ln(1 + 6L\sqrt{N})} + d\sqrt{\ln(1 + 18B^2\sqrt{d}N)} + \sqrt{\ln\frac{H}{\delta}}\right).
$$

Proof: Recall Lemma 8.4, we have with probability at least $1 - \delta$, for all n, h, for a pre-fixed V (independent of the random process):

$$
\left\|\sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)(V^\top \epsilon_h^i)\right\|^2_{(\Lambda_h^n)^{-1}} \leq 9H^2 \ln\frac{H\det(\Lambda_h^n)^{1/2}\det(\lambda I)^{-1/2}}{\delta} \leq 9H^2\left(\ln\frac{H}{\delta} + d\ln(1+N)\right)
$$

where we have used the fact that $\|\phi\|_2 \leq 1$, $\lambda = 1$, and $\|\Lambda_h^n\|_2 \leq N + 1$.

Denote the ϵ-cover of $\mathcal{F}$ as $\mathcal{N}_\epsilon$. With an application of a union bound over all functions in $\mathcal{N}_\epsilon$, we have that with probability at least $1 - \delta$, for all $V \in \mathcal{N}_\epsilon$, all n, h, we have:

$$
\left\|\sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)(V^\top \epsilon_h^i)\right\|^2_{(\Lambda_h^n)^{-1}} \leq 9H^2\left(\ln\frac{H}{\delta} + \ln(|\mathcal{N}_\epsilon|) + d\ln(1+N)\right).
$$

Recall Lemma 8.6, substitute the expression of $\ln|\mathcal{N}_\epsilon|$ into the above inequality, we get:

$$\left\|\sum_{i=1}^{n-1}\phi(s_h^i,a_h^i)(V^\top\epsilon_h^i)\right\|^2_{(\Lambda_h^n)^{-1}} \le 9H^2\left(\ln\frac{H}{\delta}+d\ln(1+6L/\epsilon)+d^2\ln(1+18B^2\sqrt{d}/\epsilon^2)+d\ln(1+N)\right).$$

Now consider an arbitrary $f\in\mathcal{F}$. By the definition of ϵ-cover, we know that for f, there exists a $V\in\mathcal{N}_\epsilon$, such that $\|f-V\|_\infty\le\epsilon$. Thus, we have:

$$\begin{aligned}
\left\|\sum_{i=1}^{n-1}\phi(s_h^i,a_h^i)(f^\top\epsilon_h^i)\right\|^2_{(\Lambda_h^n)^{-1}} &\le 2\left\|\sum_{i=1}^{n-1}\phi(s_h^i,a_h^i)(V^\top\epsilon_h^i)\right\|^2_{(\Lambda_h^n)^{-1}}+2\left\|\sum_{i=1}^{n-1}\phi(s_h^i,a_h^i)((V-f)^\top\epsilon_h^i)\right\|^2_{(\Lambda_h^n)^{-1}}\\
&\le 2\left\|\sum_{i=1}^{n-1}\phi(s_h^i,a_h^i)(V^\top\epsilon_h^i)\right\|^2_{(\Lambda_h^n)^{-1}}+8\epsilon^2N\\
&\le 9H^2\left(\ln\frac{H}{\delta}+d\ln(1+6L/\epsilon)+d^2\ln(1+18B^2\sqrt{d}/\epsilon^2)+d\ln(1+N)\right)+8\epsilon^2N,
\end{aligned}$$

where in the second inequality we use the fact that $\|\sum_{i=1}^{n-1}\phi(s_h^i,a_h^i)(V-f)^\top\epsilon_h^i\|^2_{(\Lambda_h^n)^{-1}}\le 4\epsilon^2N$, which is from

$$\left\|\sum_{i=1}^{n-1}\phi(s_h^i,a_h^i)(V-f)^\top\epsilon_h^i\right\|^2_{(\Lambda_h^n)^{-1}} \le \|\sum_{i=1}^{n-1}\phi(s_h^i,a_h^i)2\epsilon\|^2_{(\Lambda_h^n)^{-1}} \le \frac{4\epsilon^2}{\lambda}\|\sum_{i=1}^{n-1}\phi(s_h^i,a_h^i)\|_2 \le 4\epsilon^2N.$$

Set $\epsilon=1/\sqrt{N}$, we get:

$$\begin{aligned}
\left\|\sum_{i=1}^{n-1}\phi(s_h^i,a_h^i)(f^\top\epsilon_h^i)\right\|^2_{(\Lambda_h^n)^{-1}} &\le 9H^2\left(\ln\frac{H}{\delta}+d\ln(1+6L\sqrt{N})+d^2\ln(1+18B^2\sqrt{d}N)+d\ln(1+N)\right)+8\\
&\lesssim H^2\left(\ln\frac{H}{\delta}+d\ln(1+6L\sqrt{N})+d^2\ln(1+18B^2\sqrt{d}N)\right),
\end{aligned}$$

where we recall $\lesssim$ ignores absolute constants.

Now recall that we can express $\left(\widehat{P}_h^n(\cdot|s,a)-P(\cdot|s,a)\right)\cdot f=\phi(s,a)^\top(\widehat{\mu}_h^n-\mu_h^\star)^\top f$. Recall Lemma 8.3, we have:

$$\begin{aligned}
&|(\widehat{\mu}_h^n\phi(s,a)-\mu_h^\star\phi(s,a))\cdot f| \le \left|\lambda\phi(s,a)^\top(\Lambda_h^n)^{-1}(\mu_h^\star)^\top f\right|+\left|\sum_{i=1}^{n-1}\phi(s,a)^\top(\Lambda_h^n)^{-1}\phi(s_h^i,a_h^i)(\epsilon_h^i)^\top f\right|\\
&\lesssim H\sqrt{d}\,\|\phi(s,a)\|_{(\Lambda_h^n)^{-1}}+\|\phi(s,a)\|_{(\Lambda_h^n)^{-1}}\sqrt{\left(H^2\left(\ln\frac{H}{\delta}+d\ln(1+6L\sqrt{N})+d^2\ln(1+18B^2\sqrt{d}N)\right)\right)}\\
&\eqsim H\,\|\phi(s,a)\|_{(\Lambda_h^n)^{-1}}\left(\sqrt{\ln\frac{H}{\delta}}+\sqrt{d\ln(1+6L\sqrt{N})}+d\sqrt{\ln(1+18B^2\sqrt{d}N)}\right).
\end{aligned}$$

■

8.5 Algorithm

Our algorithm, Upper Confidence Bound Value Iteration (UCB-VI) will use reward bonus to ensure optimism. Specifically, we will the following reward bonus, which is motivated from the reward bonus used in linear bandit:

$$b_h^n(s,a) = \beta\sqrt{\phi(s,a)^\top(\Lambda_h^n)^{-1}\phi(s,a)}, \tag{0.7}$$

where β contains poly of H and d, and other constants and log terms. Again to gain intuition, please think about what this bonus would look like when we specialize linear MDP to tabular MDP.

Algorithm 6 UCBVI for Linear MDPs

1: **Input: parameters** β, λ
2: **for** $n = 1 \dots N$ **do**
3: Compute $\widehat{P}_h^n$ for all h (Eq. 0.3)
4: Compute reward bonus b_h^n for all h (Eq. 0.7)
5: Run Value-Iteration on $\{\widehat{P}_h^n, r_h + b_h^n\}_{h=0}^{H-1}$ (Eq. 0.8)
6: Set π^n as the returned policy of VI.
7: **end for**

With the above setup, now we describe the algorithm. Every episode n, we learn the model $\widehat{\mu}_h^n$ via ridge linear regression. We then form the quadratic reward bonus as shown in Eq. 0.7. With that, we can perform the following truncated Value Iteration (always truncate the Q function at H):

$$\begin{aligned}
&\widehat{V}_H^n(s) = 0, \forall s, \\
&\widehat{Q}_h^n(s,a) = \theta^\star \cdot \phi(s,a) + \beta\sqrt{\phi(s,a)^\top(\Lambda_h^n)^{-1}\phi(s,a)} + \phi(s,a)^\top(\widehat{\mu}_h^n)^\top \widehat{V}_{h+1}^n \\
&\qquad = \beta\sqrt{\phi(s,a)^\top(\Lambda_h^n)^{-1}\phi(s,a)} + (\theta^\star + (\widehat{\mu}_h^n)^\top \widehat{V}_{h+1}^n)^\top \phi(s,a), \\
&\widehat{V}_h^n(s) = \min\{\max_a \widehat{Q}_h^n(s,a),\ H\}, \quad \pi_h^n(s) = \operatorname{argmax}_a \widehat{Q}_h^n(s,a).
\end{aligned} \tag{0.8}$$

Note that above $\widehat{Q}_h^n$ contains two components: a quadratic component and a linear component. And $\widehat{V}_h^n$ has the format of $f_{w,\beta,\Lambda}$ defined in Eq. 0.5.

The following lemma bounds the norm of linear weights in $\widehat{Q}_h^n$.

Lemma 8.8. *Assume $\beta \in [0, B]$. For all n, h, we have $\widehat{V}_h^n$ is in the form of Eq. 0.5, and $\widehat{V}_h^n$ falls into the following class:*

$$\mathcal{V} = \{f_{w,\beta,\Lambda} : \|w\|_2 \le W + \frac{HN}{\lambda}, \beta \in [0, B], \sigma_{\min}(\Lambda) \ge \lambda\}. \tag{0.9}$$

Proof: We just need to show that $\theta^\star + (\widehat{\mu}_h^n)^\top \widehat{V}_{h+1}^n$ has its ℓ_2 norm bounded. This is easy to show as we always have $\|\widehat{V}_{h+1}^n\|_\infty \le H$ as we do truncation at Value Iteration:

$$\left\|\theta^\star + (\widehat{\mu}_h^n)^\top \widehat{V}_{h+1}^n\right\|_2 \le W + \left\|(\widehat{\mu}_h^n)^\top \widehat{V}_{h+1}^n\right\|_2.$$

Now we use the closed-form of $\widehat{\mu}_h^n$ from Eq. 0.3:

$$\left\|(\widehat{\mu}_h^n)^\top \widehat{V}_{h+1}^n\right\|_2 = \left\|\sum_{i=1}^{n-1} \widehat{V}_{h+1}^n(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top(\Lambda_h^n)^{-1}\right\|_2 \le H\left\|(\Lambda_h^n)^{-1}\sum_{i=0}^{n-1}\phi(s_h^i, a_h^i)\right\|_2 \le \frac{Hn}{\lambda},$$

where we use the fact that $\|\widehat{V}_{h+1}^n\|_\infty \le H$, $\sigma_{\max}(\Lambda^{-1}) \le 1/\lambda$, and $\sup_{s,a}\|\phi(s,a)\|_2 \le 1$. ■

8.6 Analysis of UCBVI for Linear MDPs

In this section, we prove the following regret bound for UCBVI.

Theorem 8.9 (Regret Bound). *Set $\beta = \widetilde{O}(Hd)$, $\lambda = 1$. UCBVI (Algorithm 6) achieves the following regret bound:*

$$\mathbb{E}\left[NV^\star - \sum_{i=0}^{N} V^{\pi^n}\right] \leq \widetilde{O}\left(H^2\sqrt{d^3N}\right)$$

The main steps of the proof are similar to the main steps of UCBVI in tabular MDPs. We first prove optimism via induction, and then we use optimism to upper bound per-episode regret. Finally we use simulation lemma to decompose the per-episode regret.

In this section, to make notation simple, we set $\lambda = 1$ directly.

8.6.1 Proving Optimism

Proving optimism requires us to first bound model error which we have built in the uniform convergence result shown in Lemma 8.7, namely, the bound we get for $(\widehat{P}_h^n(\cdot|s,a) - P(\cdot|s,a)) \cdot f$ for all $f \in \mathcal{V}$. Recall Lemma 8.7 but this time replacing $\mathcal{F}$ by $\mathcal{V}$ defined in Eq. 0.9. With probability at least $1-\delta$, for all n, h, s, a and for all $f \in \mathcal{V}$,

$$\begin{aligned}
\left|(\widehat{P}_h^n(\cdot|s,a) - P(\cdot|s,a)) \cdot f\right| &\leq H\left\|\phi(s,a)\right\|_{(\Lambda_h^n)^{-1}}\left(\sqrt{\ln\frac{H}{\delta}} + \sqrt{d\ln(1+6(W+HN)\sqrt{N})} + d\sqrt{\ln(1+18B^2\sqrt{d}N)}\right) \\
&\lesssim Hd\left\|\phi(s,a)\right\|_{(\Lambda_h^n)^{-1}}\left(\sqrt{\ln\frac{H}{\delta}} + \sqrt{\ln(WN+HN^2)} + \sqrt{\ln(B^2dN)}\right) \\
&\lesssim \left\|\phi(s,a)\right\|_{(\Lambda_h^n)^{-1}} \underbrace{Hd\left(\sqrt{\ln\frac{H}{\delta}} + \sqrt{\ln(W+H)} + \sqrt{\ln B} + \sqrt{\ln d} + \sqrt{\ln N}\right)}_{:=\beta}.
\end{aligned}$$

Denote the above inequality as event $\mathcal{E}_{model}$. Below we are going to condition on $\mathcal{E}_{model}$ being hold. Note that here for notation simplicity, we denote

$$\beta = Hd\left(\sqrt{\ln\frac{H}{\delta}} + \sqrt{\ln(W+H)} + \sqrt{\ln B} + \sqrt{\ln d} + \sqrt{\ln N}\right) = \widetilde{O}(Hd).$$

remark Note that in the definition of $\mathcal{V}$ (Eq. 0.9), we have $\beta \in [0, B]$. And in the above formulation of β, note that B appears inside a log term. So we need to set B such that $\beta \leq B$ and we can get the correct B by solving the inequality $\beta \leq B$ for B.

Lemma 8.10 (Optimism). *Assume event $\mathcal{E}_{model}$ is true. for all n and h,*

$$\widehat{V}_h^n(s) \geq V_h^\star(s), \forall s.$$

Proof: We consider a fixed episode n. We prove via induction. Assume that $\widehat{V}^n_{h+1}(s) \geq V^\star_{h+1}(s)$ for all s. For time step h, we have:

$$\begin{aligned}
&\hat{Q}^n_h(s,a) - Q^\star_h(s,a) \\
&= \theta^\star \cdot \phi(s,a) + \beta\sqrt{\phi(s,a)^\top(\Lambda^n_h)^{-1}\phi(s,a)} + \phi(s,a)^\top(\widehat{\mu}^n_h)^\top \widehat{V}^n_{h+1} - \theta^\star \cdot \phi(s,a) - \phi(s,a)^\top(\mu^\star_h)^\top V^\star_{h+1} \\
&\geq \beta\sqrt{\phi(s,a)^\top(\Lambda^n_h)^{-1}\phi(s,a)} + \phi(s,a)^\top \left(\widehat{\mu}^n_h - \mu^\star_h\right)^\top \widehat{V}^n_{h+1},
\end{aligned}$$

where in the last inequality we use the inductive hypothesis that $\widehat{V}^n_{h+1}(s) \geq V^\star_{h+1}(s)$, and $\mu^\star_h\phi(s,a)$ is a valid distribution (note that $\widehat{\mu}^n_h\phi(s,a)$ is not necessarily a valid distribution). We need to show that the bonus is big enough to offset the model error $\phi(s,a)^\top \left(\widehat{\mu}^n_h - \mu^\star_h\right)^\top \widehat{V}^n_{h+1}$. Since we have event $\mathcal{E}_{model}$ being true, we have that:

$$\left|(\widehat{P}^n_h(\cdot|s,a) - P(\cdot|s,a)) \cdot \widehat{V}^n_{h+1}\right| \lesssim \beta\|\phi(s,a)\|_{(\Lambda^n_h)^{-1}},$$

as by the construction of $\mathcal{V}$, we know that $\widehat{V}^n_{h+1} \in \mathcal{V}$.

This concludes the proof. ■

8.6.2 Regret Decomposition

Now we can upper bound the per-episode regret as follows:

$$V^\star - V^{\pi_n} \leq \widehat{V}^n_0(s_0) - V^{\pi_n}_0(s_0).$$

We can further bound the RHS of the above inequality using simulation lemma. Recall Eq. 0.4 that we derived in the note for tabular MDP (Chapter 7:

$$\widehat{V}^n_0(s_0) - V^{\pi_n}_0(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi_n}_h}\left[b^n_h(s,a) + \left(\widehat{P}^n_h(\cdot|s,a) - P(\cdot|s,a)\right) \cdot \widehat{V}^n_{h+1}\right].$$

(recall that the simulation lemma holds for any MDPs—it's not specialized to tabular).

In the event $\mathcal{E}_{model}$, we already know that for any s,a,h,n, we have $\left(\widehat{P}^n_h(\cdot|s,a) - P(\cdot|s,a)\right)\cdot\widehat{V}^n_{h+1} \lesssim \beta\|\phi(s,a)\|_{(\Lambda^n_h)^{-1}} = b^n_h(s,a)$. Hence, under $\mathcal{E}_{model}$, we have:

$$\widehat{V}^n_0(s_0) - V^{\pi_n}_0(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi_n}_h}\left[2b^n_h(s,a)\right] \lesssim \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi_n}_h}\left[b^n_h(s,a)\right].$$

Sum over all episodes, we have the following statement.

Lemma 8.11 (Regret Bound)**.** *Assume the event $\mathcal{E}_{model}$ holds. We have:*

$$\sum_{n=0}^{N-1}\left(V^\star_0(s_0) - V^{\pi_n}_0(s_0)\right) \leq \sum_{n=0}^{N-1}\sum_{h=0}^{H-1} \mathbb{E}_{s^n_h,a^n_h\sim d^{\pi_n}_h}\left[b^n_h(s^n_h,a^n_h)\right]$$

8.6.3 Concluding the Final Regret Bound

We first consider the following elliptical potential argument, which is similar to what we have seen in the linear bandit lecture.

Lemma 8.12 (Elliptical Potential). *Consider an arbitrary sequence of state action pairs s_h^i, a_h^i. Assume $\sup_{s,a} \|\phi(s,a)\|_2 \leq 1$. Denote $\Lambda_h^n = I + \sum_{i=0}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top$. We have:*

$$\sum_{i=0}^{N-1} \phi(s_h^i, a_h^i)(\Lambda_h^i)^{-1}\phi(s_h^i, a_h^i) \leq 2\ln\left(\frac{\det(\Lambda_h^{N+1})}{\det(I)}\right) \lesssim 2d\ln(N).$$

Proof: By the Lemma 3.7 and 3.8 in the linear bandit lecture note,

$$\begin{aligned}\sum_{i=1}^{N} \phi(s_h^i, a_h^i)(\Lambda_h^i)^{-1}\phi(s_h^i, a_h^i) &\leq 2\sum_{i=1}^{N} \ln(1 + \phi(s_h^i, a_h^i)(\Lambda_h^i)^{-1}\phi(s_h^i, a_h^i)) \\ &\leq 2\ln\left(\frac{\det(\Lambda_h^{N+1})}{\det(I)}\right) \\ &\leq 2d\ln(1 + \frac{N+1}{d\lambda}) \lesssim 2d\ln(N).\end{aligned}$$

where the first inequality uses that for $0 \leq y \leq 1$, $\ln(1+y) \geq y/2$. ■

Now we use Lemma 8.11 together with the above inequality to conclude the proof.

Proof:[Proof of main Theorem 8.9]

We split the expected regret based on the event $\mathcal{E}_{model}$.

$$\begin{aligned}\mathbb{E}\left[NV^\star - \sum_{n=1}^{N} V^{\pi_n}\right] &= \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_{model} \text{ holds}\}\left(NV^\star - \sum_{n=1}^{N} V^{\pi_n}\right)\right] \\ &\quad + \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_{model} \text{ doesn't hold}\}\left(NV^\star - \sum_{n=1}^{N} V^{\pi_n}\right)\right] \\ &\leq \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_{model} \text{ holds}\}\left(NV^\star - \sum_{n=1}^{N} V^{\pi_n}\right)\right] + \delta NH \\ &\lesssim \mathbb{E}\left[\sum_{n=1}^{N}\sum_{h=0}^{H-1} b_h^n(s_h^n, a_h^n)\right] + \delta NH.\end{aligned}$$

Note that:

$$\begin{aligned}\sum_{n=1}^{N}\sum_{h=0}^{H-1} b_h^n(s_h^n, a_h^n) &= \beta\sum_{n=1}^{N}\sum_{h=0}^{H-1}\sqrt{\phi(s_h^n, a_h^n)^\top(\Lambda_h^n)^{-1}\phi(s_h^n, a_h^n)} \\ &= \beta\sum_{h=0}^{H-1}\sum_{n=1}^{N}\sqrt{\phi(s_h^n, a_h^n)^\top(\Lambda_h^n)^{-1}\phi(s_h^n, a_h^n)} \\ &\leq \beta\sum_{h=0}^{H-1}\sqrt{N\sum_{n=1}^{N}\phi(s_h^n, a_h^n)(\Lambda_h^n)^{-1}\phi(s_h^n, a_h^n)} \lesssim \beta H\sqrt{Nd\ln(N)}.\end{aligned}$$

Recall that $\beta = \widetilde{O}(Hd)$. This concludes the proof. ■

8.7 Bibliographic Remarks and Further Readings

There are number of ways to linearly parameterize an MDP such that it permits for efficient reinforcement learning (both statistically and computationally). The first observation that such assumptions lead to statistically efficient algorithms was due to [Jiang et al., 2017] due to that these models have low Bellman rank (as we shall see in Chapter 9). The first statistically and computationally efficient algorithm for a linearly parameterized MDP model was due to [Yang and Wang, 2019a,b]. Subsequently, [Jin et al., 2020] provided a computationally and statistically efficient algorithm for simplified version of this model, which is the model we consider here. The model of [Modi et al., 2020b, Jia et al., 2020, Ayoub et al., 2020b, Zhou et al., 2020] provides another linearly parameterized model, which can viewed as parameterizing $P(s'|s,a)$ as a linear combination of feature functions $\phi(s,a,s')$. One notable aspect of the model we choose to present here, where $P_h(\cdot|s,a) = \mu_h^\star \phi(s,a)$, is that this model has a number of free parameters that is $|\mathcal{S}| \cdot d$ (note that μ is unknown and is of size $|\mathcal{S}| \cdot d$), and yet the statistical complexity does not depend on $|\mathcal{S}|$. Notably, this implies that accurate model estimation request $O(|\mathcal{S}|)$ samples, while the regret for reinforcement learning is only polynomial in d. The linearly parameterized models of [Modi et al., 2020b, Jia et al., 2020, Ayoub et al., 2020b, Zhou et al., 2020] are parameterized by $O(d)$ parameters, and, while $O(d)$ free parameters suggests lower model capacity (where accurate model based estimation requires only polynomial in d samples), these models are incomparable to the linearly parameterized models presented in this chapter;

It is worth observing that all of these models permit statistically efficient estimation due to that they have bounded Bellman rank [Jiang et al., 2017] (and bounded Witness rank [Sun et al., 2019a]), a point which we return to in the next Chapter.

The specific linear model we consider here was originally introduced by [Jin et al., 2020]. The non-parametric model-based algorithm we study here was first introduced by [Lykouris et al., 2019] (but under the context of adversarial attacks).

The analysis we present here does not easily extend to infinite dimensional feature ϕ (e.g., RBF kernel); here, [Agarwal et al., 2020a] provide an algorithm and an analysis that extends to infinite dimensional ϕ, i.e. where we have a Reproducing Kernel Hilbert Space (RKHS) and the regret is based on the concept of Information Gain.