Notes on Exploration in Linear MDPs

Nan Jiang

November 9, 2022

In this note we introduce LSVI-UCB [1] and provide its regret analysis. The note is also heavily inspired by that of Wen Sun [2]. For the purpose of clearly conveying the main ideas behind the analysis, we will be sketchy in several aspects: (1) we will omit the proof on the boundedness of the relevant parameters, especially when they appear in e.g., covering analysis in a poly-logarithmic manner, (2) we will not explicitly mention it when we need to split the failure probability over a constant number of estimation events of distinct nature. We will briefly mention it when we union bound over a variable number of events (e.g., T and H), but they will not appear in the bounds because we will use $\tilde{O}(\cdot)$ to suppress poly-logarithmic factors.

1 Setup

We consider episodic RL problems where the environment is specified by a finite-horizon MDP $M = (\mathcal{S}, \mathcal{A}, \{P_h\}, \{R_h\}, H, d_0)$. d_0 is the initial state distribution from which a trajectory is sampled and w.r.t. which the expected return $J(\pi)$ is defined, and P_h and R_h are time-dependent transition and reward functions, respectively. We adopt the convention that $r_h \in [0, 1]$, so the total sum of reward in an episode is between $[0, H]$. In finite-horizon problems, policies and value functions are also time-dependent, and we often write $\pi = \{\pi_h\}$ and $V^\pi = \{V_h^\pi\}$, where π_h and V_h^π are meant to apply to states that appear at the h-th time step.

In the **low-rank MDP** model, we have $P_h = \Phi_h \times \Psi_h \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}|}$, where $\Phi_h \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times d}$, $\Psi_h \in \mathbb{R}^{d\times|\mathcal{S}|}$, and the decomposition implies that P_h has rank at most d. The **linear MDP** setting is when the MDP has low-rank, and the matrix Φ_h is known to the learner: we denote each row of Φ_h as $\phi(s, a) \in \mathbb{R}^d$, which will be used as features by the learner. Typically, it is also assumed that the expected reward is linear in $\phi(s, a)$, i.e., $R_h(s, a) = \phi(s, a)^\top \theta_R$, and it is easy to show that (this was left as homework) $(\mathcal{T} f_{h+1})(s, a) := R_h(s, a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a))}[\max_{a'} f_{h+1}(s', a')]$ is linear in ϕ for *any* f_{h+1}; the same statement holds for $\mathcal{T}^\pi$ for any π. As a result, we have that $Q^\star$ and Q^π (for all π) are both linear in ϕ, making ϕ a powerful representation for linear MDPs.

Boundedness assumptions We assume $\|\phi(s, a)\| \leq 1\ \forall(s, a)$, where $\|\cdot\|$ without any script is the standard ℓ_2 norm. See [1] for other needed boundedness assumptions.

2 Technical Preparation

Before we introduce the algorithm and analyze it, we need to introduce a number of useful tools that find important applications across machine learning theory. Due to the purpose of this note, we only provide a very minimal introduction to some of the concepts, and will provide references for further reading on these topics.

2.1 Martingale concentration

In previous notes we have frequently used concentration inequalities for independent variables such as Hoeffding's. While they work well when data is collected by a fixed policy, the independence between data points are often broken in online exploration, as the later data points are collected using a policy computed based on previous data. (At the end of the Rmax note we already ran into such a situation.) In such cases, *Martingale concentration inequalities* become very handy. Below we introduce the simplest one, i.e., the direct extension of Hoeffding's to martingales.

Theorem 1 (Azuma's inequality). *Let $\{X_k : k = 0, 1, 2, \ldots\}$ be a martingale with $|X_k - X_{k-1}| \le c_k$ almost surely. Then, for any positive integer N, w.p. $\ge 1 - \delta$,*

$$|X_N - X_0| \le \sqrt{2\left(\sum_{k=1}^{N} c_k^2\right) \log \frac{2}{\delta}}.$$

In a typical scenario, $X_0 = 0$, and X_k is the sum of k zero-mean random variables, e.g., deviation between a random variable and its *conditional* expectation (conditioned on all previous information), and c_k is the range of the k-th r.v.. Therefore, $X_N - X_0$ is the sum of deviations across all variables, which is what we often care to bound. Different from Hoeffding's, Azuma's is more general and allows non-independent r.v.'s. This typically arises when the distribution of the random variable at time k (i.e., $X_k - X_{k-1}$) is determined based on all the information up to time $k-1$, including the realization of all previous random variables.

As an example, consider a simple multi-armed bandit. An algorithm *adaptively* pulls arms $a_1, a_2, \ldots, a_T \in \mathcal{A}$ and observes random rewards $r_1, r_2, \ldots, r_T$ for the corresponding arms, where a_t may be chosen based on $a_{1:t-1}$ and $r_{1:t-1}$. In such a case, we can still apply martingale concentration to bound $\sum_{t=1}^{T}(r_t - \mu_{a_t})$ where μ_a is the mean reward for arm a, because μ_{a_t} is the expectation of r_t given all the information so far, which includes the algorithm's choice of a_t.[1]

2.2 Ridge regression with non-stochastic inputs and the Elliptical Potential Lemma

The algorithm and analysis for linear MDPs will heavily rely on ridge regression, which we review here. Consider the following protocol: for $t = 1, 2, \ldots, T$, nature chooses $x_t \in \mathbb{R}^d$ in each round, and then reveals a noisy label $y_t = x_t^\top \theta^\star + \epsilon_t \in [0, V_{\max}]$, where ϵ_t is a zero-mean noise. $\theta^\star \in \mathbb{R}^d$ is the unknown parameters, and our goal is to learn it to make accurate predictions given any $x \in \mathbb{R}^d$, i.e., to predict $y(x) = x^\top \theta^\star$. We also assume that $\|x_t\| \le 1$ and $\|\theta^\star\|$ is a constant.

[1] In contrast, one can easily construct examples where $\sum_{t=1}^{T}(r_t - \mathbb{E}[r_t])$ does not enjoy concentration, because $\mathbb{E}[r_t]$ considers the unconditional expectation over all randomness and the partial sum does not satisfy the definition of a martingale.

When $\{x_t\}$ is sampled i.i.d. and comes with a well-conditioned *design matrix* $x_{1:T} \in \mathbb{R}^{d\times T}$, we can simply apply ordinary linear regression. However, in this setup, we allow the x_t to be chosen in an arbitrary (or even adversarial) manner, possibly depending on $x_{1:t-1}$ and $y_{1:t-1}$.

Ridge regression considers the following estimator and prediction: given $x_{1:t-1}$ and $y_{1:t-1}$,

$$\hat{\theta}_t := \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{t-1} (x_i^\top \theta - y_i)^2 + \|\theta\|^2, \qquad \hat{y}_t(x) := x^\top \hat{\theta}_t.$$

So it is simply least-square regression with ℓ_2 regularization. The estimator and the prediction can also be put in matrix form: define $\Lambda_t := I + \sum_{i=1}^{t-1} x_i x_i^\top$, we have

$$\hat{y}_t(x) = x^\top \Lambda_t^{-1} \sum_{i=1}^{t-1} x_i y_i.$$

2.2.1 "Sample complexity"

The first question we can ask is a "sample complexity" one: given data $x_{1:t-1}$ and $y_{1:t-1}$, how accurately can we make predictions about arbitrary input x? Note that we cannot expect the error to shrink to 0 when $t \to \infty$ in general, as the nature may pick a very poor design of $x_{1:t-1}$, leading to the unidentifiability of $\theta^\star$. However, even in such a case, we can still hope to make accurate predictions for x that is well within the span of $x_{1:t-1}$. This intuition is formalized by the following result:

Lemma 2 (One-shot prediction of ridge regression)**.** *For any t, w.p. $\geq 1-\delta$, $\forall x \in \mathbb{R}^d$,*

$$|\hat{y}_t(x) - y(x)| \leq \|x\|_{\Lambda_t^{-1}} \cdot O(V_{\max}\sqrt{d + \log\frac{1}{\delta}}).$$

Recall that in generalization error bounds we typically expect $1/n$ to appear on the RHS, where n is the sample size. Here, the sample size is "hidden" in Λ_t^{-1}: since Λ_t is the cumulative sum of $x_i x_i^\top$, we generally expect Λ_t to grow with the sample size. As mentioned above, however, having enough data may not suffice for accurate prediction if x falls outside the span of $x_{1:t-1}$. This is reflected in $\|x\|_{\Lambda_t^{-1}} = \sqrt{x^\top \Lambda_t^{-1} x}$, which, very roughly speaking, measures the "effective sample size", i.e., how many data points are along the direction of x.

We prove the result by quoting a vector-valued concentration inequality; see proof in [1, 2]. The proof of the concentration inequality involves martingale concentration (which we briefly introduced above) and a covering argument (which will be introduced in the next subsection).

Lemma 3. *W.p. $\geq 1-\delta$, $\|\sum_{i=1}^{t-1} x_i \epsilon_i\|_{\Lambda_t^{-1}} \leq O(V_{\max}\sqrt{d+\log\frac{1}{\delta}})$.*

Proof of Lemma 2. Under the success event of Lemma 3,

$$\begin{aligned}|\hat{y}_t(x) - y(x)| &= |x^\top \Lambda_t^{-1} \sum_{i=1}^{t-1} x_i^\top y_i - x^\top \theta^\star| \\ &\leq |x^\top \Lambda_t^{-1} \sum_{i=1}^{t-1} x_i (y_i - x_i^\top \theta^\star)| + |x^\top \Lambda_t^{-1} \sum_{i=1}^{t-1} x_i (x_i^\top \theta^\star) - x^\top \theta^\star|.\end{aligned}$$

Here we introduced an intermediate term $x^\top \Lambda_t^{-1} \sum_{i=1}^{t-1} x_i(x_i^\top \theta^\star)$, which can be interpreted as our prediction if we had observed the *noise-free* labels. By injecting this term, we bound $|\hat{y}_t(x) - y(x)|$ by the sum of two terms, where the first term can be bounded by Lemma 3. The second term is small, because the prediction based on noise-free labels would be *perfectly* accurate if (1) x is within the span of $x_{1:t-1}$, and (2) there were no regularization. As a result, bounding the second term is mainly about characterizing the bias due to regularization.

We now handle the first term:

$$
\begin{aligned}
|x^\top \Lambda_t^{-1} \sum_{i=1}^{t-1} x_i(y_i - x_i^\top \theta^\star)| &= |x^\top \Lambda_t^{-1/2} \Lambda_t^{-1/2} \sum_{i=1}^{t-1} x_i \epsilon_i| \\
\leq \|x^\top \Lambda_t^{-1/2}\| \|\Lambda_t^{-1/2} \sum_{i=1}^{t-1} x_i \epsilon_i\| &= \|x\|_{\Lambda_t^{-1}} \cdot \| \sum_{i=1}^{t-1} x_i \epsilon_i \|_{\Lambda_t^{-1}} \\
\leq \|x\|_{\Lambda_t^{-1}} \cdot O(V_{\max} \sqrt{d + \log \frac{1}{\delta}}). & \qquad \text{(Lemma 3)}
\end{aligned}
$$

For the second term,

$$
\begin{aligned}
|x^\top \Lambda_t^{-1} \sum_{i=1}^{t-1} x_i(x_i^\top \theta^\star) - x^\top \theta^\star| &= |x^\top \Lambda_t^{-1/2} \Lambda_t^{-1/2} \sum_{i=1}^{t-1} (x_i x_i^\top) \theta^\star - x^\top \Lambda_t^{-1/2} \Lambda_t^{-1/2} \Lambda_t \theta^\star| \\
= |x^\top \Lambda_t^{-1/2} \Lambda_t^{-1/2} (\sum_{i=1}^{t-1} x_i x_i^\top - \Lambda_t) \theta^\star| \|x\|_{\Lambda_t^{-1}} \|\Lambda_t^{-1/2} \cdot I \cdot \theta^\star\| & \\
\leq \|x\|_{\Lambda_t^{-1}} \|\theta^\star\|. & \qquad (\Lambda_t^{-1/2} \preccurlyeq I)
\end{aligned}
$$

Given that we assume $\|\theta^\star\|$ is a constant, this term is dominated by the first term, which completes the proof. □

2.2.2 Regret analysis

Lemma 2 shows that our accuracy of prediction relies on $\|x\|_{\Lambda_t^{-1}}$, which can be very large and may not shrink as $t \to \infty$ when the design matrix is poorly chosen. Is there a sense in which we have nice learning guarantees (i.e., some kind of error guaranteed to go to 0) for the ridge algorithm, even under arbitrary inputs?

The answer is yes, if we consider an online-learning-style protocol: in each round, after nature chooses x_t, the algorithm makes a prediction $\hat{y}_t := \hat{y}_t(x_t)$ by running ridge regression on the data observed so far, after which the true label y_t is revealed. Define the cumulative regret as

$$
\text{Regret}_T := \sum_{t=1}^{T} |\hat{y}_t - y(x_t)|.
$$

We can show that this regret is $\tilde{O}(\sqrt{T})$ (the dependence on other variables are tentatively omitted), thus the average regret $\text{Regret}_T / T \to 0$ (which can be interpreted as an average error rate) as $T \to \infty$. The intuition is that the nature can repeatedly choose the same x to form a poor design matrix, but then we will quickly learn how to predict $y(x)$ and start to incur low regret. To force a high regret, nature must choose a different direction, but we will learn as we make mistakes. Ultimately, nature will run out of directions since there are only d dimensions, which allows us to bound the total regret.

To formally bound the regret, we use Lemma 2 and union bound over T steps (which incurs an $O(\log T)$ dependence suppressed by $\tilde{O}(\cdot)$ notation): w.p. $\geq 1-\delta$,

$$\sum_{t=1}^{T} |\hat{y}_t - y(x_t)| \leq \sum_{t=1}^{T} \|x_t\|_{\Lambda_t^{-1}} \cdot \tilde{O}(V_{\max}\sqrt{d+\log\frac{1}{\delta}}) \tag{1}$$

$$\leq \tilde{O}(V_{\max}\sqrt{d+\log\frac{1}{\delta}}) \cdot \sqrt{T}\sqrt{\sum_{t=1}^{T} x_t^\top \Lambda_t^{-1} x_t} \qquad \text{(Jensen's)}$$

Since we already have $\sqrt{T}$, it remains to bound $\sum_{t=1}^T x_t^\top \Lambda_t^{-1} x_t$ in a way that does not incur further polynomial dependence on T, which is given by the famous elliptical potential lemma. Note that $\sum_{t=1}^T x_t^\top \Lambda_t^{-1} x_t$ only depends on $\{x_t\}_{t=1}^T$ which is chosen arbitrarily, so the elliptical potential lemma is a general result that does not require any assumptions other than $\|x_t\| \leq 1\ \forall t$.

Lemma 4 (Elliptical potential lemma)**.**

$$\sum_{t=1}^{T} x_t^\top \Lambda_t^{-1} x_t \leq 2\log\det(\Lambda_{T+1}) \leq 2d\log(T+1).$$

Proof. First, $\forall t$, because $\Lambda_t^{-1} \preccurlyeq I$, $x_t^\top \Lambda_t^{-1} x_t \leq 1$. Then, using the fact that $z \leq 2\log(1+z)$ for any $z \in [0,1]$, we have $x_t^\top \Lambda_t^{-1} x_t \leq 2\log(1 + x_t^\top \Lambda_t^{-1} x_t)$.

On the other hand, let $\Lambda_0 := I$, and for all $t \geq 0$,

$$\begin{aligned} \det(\Lambda_{t+1}) &= \det(\Lambda_t + x_t x_t^\top) \\ &= \det(\Lambda_t^{1/2}(I + \Lambda_t^{-1/2} x_t x_t^\top \Lambda_t^{-1/2})\Lambda_t^{1/2}) \\ &= \det(\Lambda_t)\det(I + \Lambda_t^{-1/2} x_t x_t^\top \Lambda_t^{-1/2}) \\ &= \det(\Lambda_t)\cdot(1 + x_t \Lambda_t^{-1} x_t^\top). \end{aligned}$$

For the last step, note that for $A \in \mathbb{R}^{m\times n}$, $A^\top A$ and $AA^\top$ share the same non-zero eigenvalues, so $\det(I_n + A^\top A) = \det(I_m + AA^\top)$. The step follows from letting $A = \Lambda_t^{-1/2} x_t \in \mathbb{R}^{d\times 1}$.

Now, taking log on both sides, we have $\log(\det(\Lambda_{t+1})/\det(\Lambda_t)) = \log(1 + x_t\Lambda_t^{-1}x_t^\top) \geq \frac{1}{2} x_t \Lambda_t^{-1} x_t^\top$, so

$$\sum_{t=1}^{T} x_t^\top \Lambda_t^{-1} x_t \leq 2\sum_{t=1}^{T} \log\frac{\det(\Lambda_{t+1})}{\det(\Lambda_t)} = 2\log\frac{\det(\Lambda_{T+1})}{\det(\Lambda_0)}.$$

This proves the first inequality given $\Lambda_0 = I$.

To further bound $\det(\Lambda_{T+1})$, we have $\det(A) \leq \sigma_{\max}(A)^d$ where $\sigma_{\max}(\cdot)$ is the largest eigenvalue.

$$\sigma_{\max}(\Lambda_{T+1}) = \max_{\|u\|=1} u^\top(I + \sum_{t=1}^{T} x_t x_t^\top)u \leq T+1. \qquad (\|x_t\| \leq 1) \quad \square$$

2.3 Covering

In previous lectures we have analyzed function-approximation algorithms, assuming a finite function class $\mathcal{F}$. We simply union bound over it in generalization analyses and pay $\log|\mathcal{F}|$. However, almost all function classes in practice (including linear classes) are continuous, in which case the cardinality of the function class is infinite. The covering argument provides a way to derive generalization error bounds for infinite classes. Given a (possibly infinite) function class $\mathcal{F} \subset (\mathcal{X} \to \mathbb{R})$, its ℓ_∞-covering number is defined as

Definition 1 (ℓ_∞ covering number)**.** $\mathcal{F}_\epsilon \subset (\mathcal{X} \to \mathbb{R})$ is an ϵ-cover of $\mathcal{F}$, if $\forall f \in \mathcal{F}$, $\exists f_C \in \mathcal{F}_\epsilon$ such that $\|f - f_C\|_\infty \leq \epsilon$. The ℓ_∞ covering number $\mathcal{N}_\epsilon := \mathcal{N}_\epsilon(\mathcal{F})$ is the size of the smallest ϵ-cover.

2.3.1 Using covering number in generalization bounds

We first show how to prove generalization bounds for $\mathcal{F}$ when it admits a finite covering number. Consider a simple setting: we have $x_1, \ldots, x_n \in \mathcal{X}$ sampled i.i.d. from some distribution, and want to bound $\sup_{f\in\mathcal{F}} |\mathbb{E}[f(X)] - \frac{1}{n}\sum_{i=1}^n f(x_i)|$. Let's assume that $f \in [0, 1]$.

Let $\mathcal{F}_\epsilon$ be the (smallest) ϵ-cover of $\mathcal{F}$, where the value of ϵ will be decided later. Then, by Hoeffding + union bound, w.p. $\geq 1 - \delta$, $\forall f_C \in \mathcal{F}_\epsilon$, $|\mathbb{E}[f_C(X)] - \frac{1}{n}\sum_{i=1}^n f_C(x_i)| \leq \sqrt{\frac{1}{2n}\log\frac{2\mathcal{N}_\epsilon}{\delta}}$. Note that this will be the only high-probability statement we make.

We now try to use the above to bound $|\mathbb{E}[f(X)] - \frac{1}{n}\sum_{i=1}^n f(x_i)|$. Fixing any $f \in \mathcal{F}$, find its closest cover center $f_C \in \mathcal{F}_\epsilon$ with $\|f - f_C\|_\infty \leq \epsilon$, then

$$\begin{aligned}|\mathbb{E}[f(X)] - \frac{1}{n}\sum_{i=1}^n f(x_i)| \leq{} & |\mathbb{E}[f_C(X)] - \frac{1}{n}\sum_{i=1}^n f_C(x_i)| \\ & + |\mathbb{E}[f(X)] - \mathbb{E}[f_C(X)]| \\ & + |\tfrac{1}{n}\textstyle\sum_{i=1}^n f(x_i) - \tfrac{1}{n}\sum_{i=1}^n f_C(x_i)|.\end{aligned}$$

Here the first term is bounded by union bounding over $\mathcal{F}_\epsilon$, and the second and the third terms are each *deterministically* bounded by ϵ due to $\|f - f_C\|_\infty \leq \epsilon$. So overall the generalization error bound is $2\epsilon + \sqrt{\frac{1}{2n}\log\frac{2\mathcal{N}_\epsilon}{\delta}}$, and one can optimize ϵ to obtain the best bound.

As we will see below, typically $\log \mathcal{N}_\epsilon$ grows with ϵ very slowly, often as $\log\frac{1}{\epsilon}$. Therefore, as long as we do not care about poly-logarithmic terms, ϵ can be set arbitrarily small as long as it is inversely polynomial, and as long as ϵ is sufficiently small, $\sqrt{\frac{1}{2n}\log\frac{2\mathcal{N}_\epsilon}{\delta}}$ can be roughly viewed as not changing with ϵ (since the change is only poly-logarithmic). Typically, one can set $\epsilon = \sqrt{\frac{1}{2n}\log\frac{2\mathcal{N}_\epsilon}{\delta}}$, since further reducing ϵ can at most improve the bound by a constant factor of 3.

2.3.2 Bounding the covering number of a linear class

The other half of the puzzle is how to bound the covering number for common classes. Here we show a simple example of linear class: consider $\mathcal{F} = \{x \mapsto \phi(x)^\top\theta : \|\theta\| \leq B\}$, where the feature map ϕ satisfies $\|\phi(x)\| \leq 1\ \forall x$. We want to bound $\mathcal{N}_\epsilon(\mathcal{F})$.

We construct $\mathcal{F}_\epsilon$ as follows: first, $\|\theta\| \leq B$ implies that $\|\theta\|_\infty \leq B$, so we can relax the ℓ_2 ball of θ to a larger ℓ_∞ ball for easy construction. Then, we simply create a regular grid of resolution ϵ' for $\{\theta : \|\theta\|_\infty \leq B\}$, which results in $(B/\epsilon')^d$ grids[2]. Let θ_C be the center of a grid, and any other θ in the same grid we have $\|\theta - \theta_C\| \leq \sqrt{d}\|\theta - \theta_C\|_\infty \leq \sqrt{d}\epsilon'$. Choosing the functions correspond to the grid centers immediately yield an ϵ-cover of $\mathcal{F}$, with $\|f - f_C\|_\infty = \max_x |\phi(x)^\top(\theta - \theta^C)| \leq \max_x \|\phi(x)\|\|\theta - \theta^C\| \leq \sqrt{d}\epsilon'$. So to obtain an ϵ-cover, we can set $\epsilon' = \epsilon/\sqrt{d}$, and $\mathcal{N}_\epsilon \leq (B\sqrt{d}/\epsilon)^d$.

[2] We ignore the issue that B/ϵ' may not be an integer; this can be easily handled by relaxing the cover size to $(2B/\epsilon')^d$.

Recall that we pay the log-covering number in the generalization error bounds, which is $\log \mathcal{N}_\epsilon \leq d \log \frac{B\sqrt{d}}{\epsilon}$. So, as long as B and ϵ are polynomial, they only incur poly-logarithmic dependence in $\log \mathcal{N}_\epsilon$, and the main part of $\log \mathcal{N}_\epsilon$ is simply d.

Remark When we construct the cover, we created grids which can be viewed as ℓ_∞ "balls" in the parameter space $\mathbb{R}^d$. This ℓ_∞ should not be confused with the "ℓ_∞" in ℓ_∞ covering, as the latter refers to the fact that we measure the closeness between two functions as $\|f - f'\|_\infty = \max_x |f(x) - f'(x)|$, where the ℓ_∞ corresponds to taking max over all possible function inputs. In fact, since we later relaxed $\|\theta - \theta_C\|_\infty$ to $\|\theta - \theta_C\|$, it is also valid to say that we essentially created a ℓ_2 cover over the parameter space, which nevertheless led to an ℓ_∞ cover over the function space.

Besides ℓ_∞ covering (which is the simplest), there are alternatives such as ℓ_1 covering which measures the distance between two functions in a *sample-dependent* manner, e.g., $\frac{1}{n}\sum_{i=1}^n |f(x_i) - f'(x_i)|$. When using such a covering number in generalization analysis, one cannot reduce to the analysis to the case of finite classes due to the sample dependence, and the analysis requires techniques such as symmetrization similar to what is used in the analysis of VC dimensions; see [3, 4].

2.3.3 Lipschitz composition

In Section 2.3.1 we show how to bound $\sup_{f \in \mathcal{F}} |\mathbb{E}[f(X)] - \frac{1}{n}\sum_{i=1}^n f(x_i)|$ when $\mathcal{F}$ admits an ℓ_∞ covering number. In actual analyses, we rarely directly care about the average of f itself, but often some loss function over f, e.g., $\frac{1}{n}\sum_{i=1}^n |y_i - f(x_i)|$. How to handle such cases?

We consider a more general setup where we want to bound $|\mathbb{E}[l(\cdot\,; f)] - \hat{\mathbb{E}}[l(\cdot\,; f)]|$ for all $f \in \mathcal{F}$, where $(\cdot)$ is a placeholder for the data which can include more than x_i (e.g., it is (x_i, y_i) in the above example), and $\hat{\mathbb{E}}$ is the empirical average. The overall idea is still similar: we consider $\mathcal{F}_\epsilon$ (the cover of $\mathcal{F}$), and decompose the difference as:

$$\begin{aligned} |\mathbb{E}[l(\cdot\,; f)] - \hat{\mathbb{E}}[l(\cdot\,; f)]| \leq\ & |\mathbb{E}[l(\cdot\,; f_C)] - \hat{\mathbb{E}}[l(\cdot\,; f_C)]| \\ & + |\mathbb{E}[l(\cdot\,; f)] - \mathbb{E}[l(\cdot\,; f_C)]| \\ & + |\hat{\mathbb{E}}[l(\cdot\,; f)] - \hat{\mathbb{E}}[l(\cdot\,; f_C)]|. \end{aligned}$$

As before, the first term can be controlled by union bounding over $\mathcal{F}_\epsilon$, and the remaining terms by $\|f - f_C\|_\infty$. However, there are two notable changes:

- For the first term, it is the range of l that matters in the concentration bounds, not f.
- When controlling the remaining terms, we typically need l to be Lipcshitz in f, in the sense that

  $$|l(\cdot\,; f) - l(\cdot\,; f_C)| \leq L\|f - f_C\|_\infty.$$

  To make sure the two terms are controlled by ϵ (which is often set to be equal to the concentration bound for the first term), we will need an (ϵ/L)-cover of $\mathcal{F}$, so the Lipschitz constant L enters the covering number. Fortunately, it will typically only appear poly-logarithmically in the log-covering number, so as long as L is polynomial we are fine.

Another way to look at this is that we are essentially trying to translate the cover of $\mathcal{F}$ to a cover of another function class $\{l(\cdot\,; f) : f \in \mathcal{F}\}$, and via doing so we incur dependence on the Lipschitzness of l in f.

3 Algorithm

We are now ready to state the algorithm. Let (s_h^i, a_h^i, r_h^i) denote the state-action-reward tuple encountered in the i-th episode. For each episode $t = 1, 2, \ldots, T$, the algorithm performs an optimistic version of LSVI (least-square value iteration): for $h = H, H-1, \ldots, 1$:

1. Define $\Lambda_t^h := I + \sum_{i=1}^{t-1} \phi_h^i (\phi_h^i)^\top$, where $\phi_h^i := \phi(s_h^i, a_h^i)$.
2. Let $\widetilde{Q}_h^t$ be the point estimator of ridge regression:

$$\widetilde{Q}_h^t(s,a) := \phi(s,a)^\top (\Lambda_h^t)^{-1} \sum_{i=1}^{t-1} \phi_h^i (r_h^i + \widehat{V}_{h+1}(s_{h+1})). \tag{2}$$

3. Add bonus to ensure optimism: $\widehat{Q}(s,a) := \widetilde{Q}(s,a) + \beta\sqrt{\phi(s,a)^\top (\Lambda_h^t)^{-1} \phi(s,a)}$.
4. Clipping: $\widehat{V}_h^t(s) := \min\{H, \max_a \widehat{Q}_h^t(s,a)\}$.

We adopt the convention that all notions of value functions evaluate to 0 at $h = H+1$. $\beta > 0$ is a hyperparameter which will be set in the analysis. Once the computation is completed at $h = 1$, the algorithm collects a new episode of data $(s_1^t, a_1^t, r_1^t, \ldots, s_H^t, a_H^t, r_H^t)$ using π^t, the greedy policy w.r.t. $\{\widehat{Q}_h^t\}$.

4 Regret analysis

4.1 Concentration bounds

We establish concentration bounds first, and then use them to bound the regret of the algorithm. A key step of the algorithm is to run ridge regression at each level h, where x_i corresponds to $\phi(s_h^i, a_h^i)$ and y_i corresponds to $r_h^i + V(s_{h+1})$ for $V = \widehat{V}_{h+1}$. While we would be able to directly apply the guarantee for ridge regression in Section 2.2.1 if V is fixed a priori, the data-dependence of $\widehat{V}_{h+1}$ violates the independence and invalidates the guarantees.

Remark: Bellman-completeness is insufficient for LSVI-UCB Before diving into how the issue of data-dependence of $\widehat{V}_{h+1}$, we emphasize that the above application of ridge regression crucially relies on the fact that the expected label, $R_h(s,a) + P_h(s,a)^\top V$ is linear in ϕ *regardless of the choice of* V. Recall that in the FQI section we have seen the completeness assumption $\mathcal{T} f \in \mathcal{F} \forall f \in \mathcal{F}$ (or the $\mathcal{T}^\pi$ version), and in linear MDPs the linear class (in ϕ) satisfies both assumptions. However, the algorithm would not work if we merely assume $\mathcal{F}$ is linear and Bellman-complete (instead of the MDP being a linear MDP), because it needs to back up functions *outside* the linear class: while $\widetilde{Q}$ is linear, we add a (square-root of) quadratic bonus to it, followed by clipping, both of which makes the Q-function we backup nonlinear. On the other hand, when only linear completeness is assumed, there are information-theoretic algorithms that can handle it, e.g., by showing that the problem has low "Q-type" Bellman rank [5, Sec 9.3.1].

To handle the data-dependence of $\widehat{V}_{h+1}$, we identify a function class that is defined independent of the data and is guaranteed to contain $\widehat{V}_{h+1}$, and try to union bound over it. The class is

$$\mathcal{V} := \{V_{w,\Lambda}, \|w\| \leq B, \sigma_{\min}(\Lambda) \geq 1\},$$

where $V_{w,\Lambda}(s) := \min\{H, \max_a(w^\top \phi(s,a) + \beta\sqrt{\phi(s,a)^\top \Lambda^{-1}\phi(s,a)})\}$. It is easy to see that $\widehat{V}_{h+1} \in \mathcal{V}$ as long as B is a deterministic upper bound of the linear coefficient in Eq.(2) (i.e., the part of RHS after $\phi(s,a)^\top$). To obtain this bound, note that for each i, $\|\phi_h^i(r_h^i + \widehat{V}_{h+1}(s_{h+1}))\| \leq (H+1)$, so[3]

$$\|(\Lambda_h^t)^{-1}\sum_{i=1}^{t-1}\phi_h^i(r_h^i + \widehat{V}_{h+1}(s_{h+1}))\| \leq \sum_{i=1}^{t-1}\|\phi_h^i(r_h^i + \widehat{V}_{h+1}(s_{h+1}))\| \leq tH(H+1).$$

Apparently we cannot directly union bound over $\mathcal{V}$ since it is a continuous class, and instead we use a covering argument. The result is that we can obtain an α-cover of $\mathcal{V}$ of size $\tilde{O}(d^2)$ for any polynomially small α. We will only give a proof sketch here: the main idea is to create covers for "simple" classes (i.e., linear and quadratic), and then analyze the effects of the transformations .

1. First, create an $\alpha/2$-cover of $\{\phi^\top w : \|w\| \leq B\}$ and an $\alpha^2/4$ cover of $\{\phi^\top\Lambda^{-1}\phi^\top : \sigma_{\min}(\Lambda) \geq 1\}$, with log cardinalities $\tilde{O}(d)$ and $\tilde{O}(d^2)$, respectively. Note that the latter is effectively a linear class with features $\{\phi_i\phi_j\}_{i,j}$, and the linear coefficients are the entries of Λ^{-1} which are bounded due to $\sigma_{\min}(\Lambda) \geq 1$.
2. Let $\{\Lambda_C\}$ be the parameters of the cover centers of the quadratic class. We show that these parameters also provide an $\alpha/2$ cover for the square-root class $\{\sqrt{\phi^\top\Lambda^{-1}\phi^\top} : \sigma_{\min}(\Lambda) \geq 1\}$, because for any Λ and its closest Λ_C, $|\sqrt{\phi^\top\Lambda^{-1}\phi^\top} - \sqrt{\phi^\top\Lambda_C^{-1}\phi^\top}| \leq |\sqrt{\phi^\top\Lambda^{-1}\phi^\top - \phi^\top\Lambda_C^{-1}\phi^\top}| \leq \alpha/2$. The first inequality follows from the fact that $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x-y|}, \forall x, y \geq 0$.
3. The additive composition of the linear and the square-root class yields an α-cover.
4. $\min\{H, \cdot\}$ and $\max_a(\cdot)$ are non-expansions in $\|\cdot\|_\infty$ (recall $|\max_a f(s,a) - \max_a f'(s,a)| \leq \max_a |f(s,a) - f'(s,a)|$), so we have an α-cover of the final class, whose size is the product of the covering numbers for the linear and the quadratic classes. The log covering number is then dominated by that of the quadratic class, which is $\tilde{O}(d^2)$.

Let $\mathcal{V}_\alpha$ be the resulting α-cover of $\mathcal{V}$. For any fixed t and h, we now apply Lemma 2 for any fixed V, with the following correspondence between the variables:

- $x = \phi(s,a)$ (which we will abbreviate as ϕ) is the queried input.
- $y(x) = R_h(s,a) + P_h(s,a)^\top V$.
- $\hat{y}_t(x) = \widetilde{Q}_h^t(s,a;V) := \phi(s,a)^\top(\Lambda_h^t)^{-1}\sum_{i=1}^{t-1}\phi_h^i(r_h^i + V(s_{h+1}))$.
- $\Lambda_t = \Lambda_h^t$, $V_{\max} = H+1$.

Union bounding over all $V = V_C \in \mathcal{V}_\alpha$, we obtain that w.p. $\geq 1-\delta$, $\forall V_C \in \mathcal{V}_\alpha$, $\forall(s,a)$,

$$|\widetilde{Q}_h^t(s,a;V_C) - R_h(s,a) - P_h(s,a)^\top V_C| \leq \|\phi\|_{(\Lambda_h^t)^{-1}}\tilde{O}(H(d + \sqrt{\log\tfrac{1}{\delta}})).$$

[3] See [1, Lemma B.2] for a bound that is tighter in t but incurs dependence on d.

Recall our goal is to bound the LHS of the above for all $V \in \mathcal{V}$, and we use an argument similar to Section 2.3.3: for any $V \in \mathcal{V}$, we decompose the error as follows:

$$\begin{aligned}|\widetilde{Q}_h^t(s,a;V) - R_h(s,a) - P_h(s,a)^\top V| \le & |\widetilde{Q}_h^t(s,a;V_C) - R_h(s,a) - P_h(s,a)^\top V_C| \\ & + |\widetilde{Q}_h^t(s,a;V_C) - \widetilde{Q}_h^t(s,a;V)| \\ & + |P_h(s,a)^\top V - P_h(s,a)^\top V_C|.\end{aligned}$$

Here the first term is handled by union bounding over $\mathcal{V}_\alpha$, and recall from Section 2.3.3 that it suffices to show that we can use α to control the remaining terms up to polynomial blow-up factors. The third term is easily bounded by α, so the key is to bound the second term: let $\Delta_i := V(s_{h+1}^i) - V_C(s_{h+1})^i$

$$|\widetilde{Q}_h^t(s,a;V_C) - \widetilde{Q}_h^t(s,a;V)| = |\phi^\top(\Lambda_h^t)^{-1}\sum_{i=1}^{t-1}\phi_h^i\Delta_i| \le \|\phi\|\|\sum_{i=1}^{t-1}\phi_h^i\Delta_i\| \le (t-1)\alpha.$$

Having verified that the blow-up factors are polynomial, we have w.p. $\ge 1-\delta$, for all $t \in [T]$ and $h \in [H]$ (the additional $\log(HT)$ dependence is suppressed by $\tilde{O}(\cdot)$): define the residual of ridge regression as

$$b_h^t(s,a) := |\widetilde{Q}_h^t(s,a) - R_h(s,a) - P_h(s,a)^\top \widehat{V}_{h+1}^t|, \tag{3}$$

then

$$b_h^t(s,a) \le \|\phi\|_{(\Lambda_h^t)^{-1}}\tilde{O}(H(d+\sqrt{\log\tfrac{1}{\delta}})) =: \beta\|\phi\|_{(\Lambda_h^t)^{-1}}. \tag{4}$$

Here we choose the hyperparameter of the algorithm β as the coefficient $\tilde{O}(H(d+\sqrt{\log\frac{1}{\delta}}))$ in the bound. Since $\widehat{Q}$ is defined as $\widetilde{Q} + \beta\|\phi\|_{(\Lambda_h^t)^{-1}}$, a direct consequence is

$$\widehat{Q}_h^t \ge R_h + P_h^\top \widehat{V}_{h+1}^t. \tag{5}$$

4.2 Regret bound

Before bounding the regret, we first prove that the algorithm is indeed optimistic. All results in this section are based on the success of the event in Eq.(4).

Lemma 5 (Optimism). *For all h, t, $\widehat{Q}_h^t \ge Q_h^*$, $\widehat{V}_h^t \ge V_h^*$.*

Proof. The base case $h = H+1$ is trivial as all value functions are 0. Inductively:

$$\begin{aligned}\widehat{Q}_h^t(s,a) &\ge R_h(s,a) + P_h(s,a)^\top \widehat{V}_{h+1}^t && \text{(Eq.(5))} \\ &\ge R_h(s,a) + P_h(s,a)^\top V_{h+1}^* && \text{(Inductive hypothesis for } \widehat{V}_{h+1}^t) \\ &= Q_h^*(s,a).\end{aligned}$$

To show $\widehat{V}_h^t \ge V_h^\star$, note that $\max_a \widehat{Q}_h^t(s,a) \ge \max_a Q_h^\star(s,a) = V^\star(s)$, and clipping at H does not matter because $V_h^\star \le H$. □

With this, we are finally ready to analyze the regret of the algorithm:

$$
\begin{aligned}
\text{Regret}_T &:= \sum_{t=1}^{T} J(\pi^\star) - J(\pi^t) \\
&\le \sum_{t=1}^{T} \mathbb{E}_{s_1\sim d_0}[\widehat{Q}_1^t(s_1, \pi^t)] - J(\pi^t) && \text{(Optimism)} \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi^t}}[\widehat{Q}_h^t(s_h, a_h) - R_h(s_h, a_h) - P_h(s_h, a_h)^\top[\max_a \widehat{Q}_h^t(\cdot, a_{h+1})]] \\
&\le \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi^t}}[\widehat{Q}_h^t(s_h, a_h) - R_h(s_h, a_h) - P_h(s_h, a_h)^\top[\min\{H, \max_a \widehat{Q}_h^t(\cdot, a_{h+1})]\}] \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi^t}}[\widetilde{Q}_h^t(s_h, a_h) + \beta\|\phi(s_h, a_h)\|_{(\Lambda_h^t)^{-1}} - R_h(s_h, a_h) - P_h(s_h, a_h)^\top \widehat{V}_{h+1}] \\
&\le \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi^t}}[b_h^t(s_h, a_h) + \beta\|\phi(s_h, a_h)\|_{(\Lambda_h^t)^{-1}}] && \text{(See Eq.(3) for the def of } b_h^t\text{)} \\
&\le \sum_{h=1}^{H}\sum_{t=1}^{T} \mathbb{E}_{d_h^{\pi^t}}[2\beta\|\phi(s_h, a_h)\|_{(\Lambda_h^t)^{-1}}] && \text{(Eq.(4))}
\end{aligned}
$$

The third line is essentially the finite-horizon variant of a telescoping lemma we have used multiple times in previous lectures, where we translate the error of using an arbitrary Q to evaluate $J(\pi)$ into the Bellman error of Q along the occupancy of π.

We are almost there—what we obtain above looks very similar to (an upper bound of) the cumulative regret of ridge regression (Eq.(1)). However, there is still a crucial gap: in Eq.(1), we can upper-bound the sum of $\|x_t\|_{\Lambda_t^{-1}}$ for those x_t that goes into the definition of Λ_t. If we want to apply this bound, we need to have $\|\phi(s_h^t, a_h^t)\|_{(\Lambda_h^t)^{-1}}$ show up, where s_h^t, a_h^t is the actual state-action pair we run into in the t-th episode. Instead, what we have is an expectation over $(s_h, a_h) \sim d_h^{\pi^t}$.

Fortunately, this mismatch can be addressed by noting that (1) s_h^t, a_h^t is actually sampled from $d_h^{\pi^t}$, (2) $2\beta\|\phi(s_h, a_h)\|_{(\Lambda_h^t)^{-1}}$ as a function $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined deterministically based on all the information before episode t, and (3) the function is bounded between $[0, \beta]$. So, we can directly apply Azuma's inequality to bridge the gap between $(s_h, a_h) \sim d_h^{\pi^t}$ and s_h^t, a_h^t, and the remaining steps use Jensen and the elliptical potential lemma in the same way as Section 2.2.2

$$
\begin{aligned}
\text{Regret}_T &\le \sum_{h=1}^{H}\sum_{t=1}^{T} \mathbb{E}_{d_h^{\pi^t}}[2\beta\|\phi(s_h, a_h)\|_{(\Lambda_h^t)^{-1}}] \\
&\le \sum_{h=1}^{H}\sum_{t=1}^{T} 2\beta\|\phi(s_h^t, a_h^t)\|_{(\Lambda_h^t)^{-1}} + \tilde{O}(\beta\sqrt{T\log\tfrac{1}{\delta}}) \\
&\le 2\beta\sum_{h=1}^{H}\sqrt{T}\sum_{t=1}^{T}\phi_h^t(\Lambda_h^t)^{-1}\phi_h^t + \tilde{O}(\beta\sqrt{T\log\tfrac{1}{\delta}}) && \text{(Jensen)} \\
&\le 2\beta H\sqrt{T}\sqrt{2d\log(T+1)} + \tilde{O}(\beta\sqrt{T\log\tfrac{1}{\delta}}) && \text{(Elliptical potential lemma)} \\
&= \tilde{O}(H^2\sqrt{d}(d + \sqrt{\log\frac{1}{\delta}})\sqrt{T}). && \text{(Plug in def of } \beta \text{ from Eq.(4))}
\end{aligned}
$$

.

References

[1] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[2] Wen Sun. *Learning in Linear MDPs: Upper Confidence Bound Value Iteration*. `https://wensun.github.io/CS6789_data/linearMDP_note_complete.pdf`.

[3] Sham Kakade and Ambuj Tewari. *CMSC 35900 (Spring 2008) Learning Theory: Covering Numbers*. `https://home.ttic.edu/~tewari/lectures/lecture14.pdf`.

[4] Shivani Agarwal. *E0 370 Statistical Learning Theory: Covering Numbers, Pseudo-Dimension, and Fat-Shattering Dimension*. `http://www.shivani-agarwal.net/Teaching/E0370/Aug-2011/Lectures/5.pdf`.

[5] Alekh Agarwal, Nan Jiang, Sham Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms*. `https://rltheorybook.github.io/`.