CS 839 Probability and Learning in High Dimension Lecture 22 - 04/20/2022

Lecture 22: Linear MDPs II

Lecturer: Yudong Chen *Scribe: Ting Cai*

In this lecture,[1] we recap the structure of Linear MDPs and the episodic setting. We then introduce the *Least-Squares Value Iteration with Upper Confidence Bound* (LSVI-UCB) algorithm and the regret bound for this algorithm. We also discuss the proof of the regret bound.

1 Recap: Linear structure

We assume that both the reward function and transition kernel have a linear structure, with respect to some *known* feature map. In the sequel $\|\cdot\|$ denotes the ℓ_2 norm on $\mathbb{R}^d$.

Assumption 1 (Linearity and Boundedness)**.** *For each $h \in [H]$ and $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, it holds that*

$$\mathbb{P}_h(x'|x,a) = \langle \phi(x,a), \mu_h(x') \rangle \qquad \text{and}$$
$$r_h(x,a) = \langle \phi(x,a), \theta_h \rangle,$$

where

- $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ *is a* ***known*** *feature map,*
- $\mu_h = (\mu_h^{(i)})_{i \in [d]}$ *is a vector of d* ***unknown*** *(signed) measures on $\mathcal{S}$, and*
- $\theta_h = (\theta_h^1, \ldots, \theta_h^d) \in \mathbb{R}^d$ *is a vector of d* ***unknown*** *weights.*

We assume that $\max_{(x,a)\in\mathcal{S}\times\mathcal{A}} \|\phi_h(x,a)\| \le 1$, $\|\mu_h(\mathcal{S})\| \le \sqrt{d}, \|\theta_h\| \le \sqrt{d}$ *for all $h \in [H]$.*[2]

The above assumption implies that the Q-function is linear for any policy (including the optimal policy).

Lemma 1 (Linearity of Q)**.** *For any policy π and $h \in [H]$, there exists a weight vector $w_h^\pi \in \mathbb{R}^d$ such that*

$$Q_h^\pi(x,a) = \langle \phi(x,a), w_h^\pi \rangle, \qquad \forall (x,a) \in \mathcal{S} \times \mathcal{A}.$$

In particular, the optimal Q-function satisfies $Q_h^*(x,a) = \langle \phi(x,a), w_h^* \rangle, \forall x, a$ for some $w_h^* \in \mathbb{R}^d$.

2 Episodic setting and regrets

The agent interacts with the MDP in K episodes. At the beginning of episode k, the agent picks a policy $\pi^k = (\pi_1^k, \ldots, \pi_H^k)$ and receives an (arbitrary) initial state x_1^k. The agents then executes the policy for H steps, resulting in the trajectory

$$x_1^k, a_1^k, r_1^k, \ldots, x_H^k, a_H^k, r_H^k,$$

where $a_h^k \sim \pi_h^k(x_h^k)$, $r_h^k = r(x_h^k, a_h^k)$ and $x_{h+1}^k \sim \mathbb{P}_h(\cdot|x_h^k, a_h^k)$. The system then resets and episode $(k+1)$ begins.

The regret over K episodes is defined as

$$\text{Regret}(K) := \sum_{k=1}^K \left[V_1^*(x_1^k) - V_1^{\pi^k}(x_1^k) \right],$$

which is the difference between the total value of the agent's policy $\pi^1, \ldots, \pi^K$ and that of the optimal policy. We want to find an algorithm that achieves a low regret.

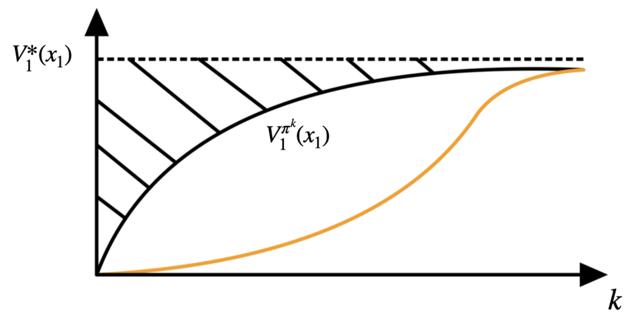

Figure 1: Illustration of value functions across episodes. The dashed line represents $V_1^*(x_1)$, indicating the highest possible return one can get in each episode. The black solid line represents the value function as the episode k progresses. The yellow line represents the value function by another possible algorithm, which has higher regret.

Suppose all K episodes start at the same initial state x_1. The regret correspond to the shaded area in Figure 1, which we aim to minimize.

Remark 2. In Figure 1, both algorithms (black curve and yellow curve) eventually converge to the optimal value when k is sufficiently large. However, the algorithm corresponding to the black curve has smaller regret *during the learning process*.

Remark 3. Suppose the (total) regret grows sublinearly, i.e., $\text{Regret}(K) = o(K)$. In this case, the average regret $\text{AvgRegret}(K) := \frac{1}{K}\sum_{k=1}^{K}\left[V_1^*(x_1^k) - V_1^{\pi^k}(x_1^k)\right] = o(1)$ ultimately goes to 0. For $\text{AvgRegret}(K)$, it's possible for a few episodes to have very bad value functions, but the effect of the bad episodes won't matter as it will eventually average out.

3 Algorithm and guarantees

The algorithm, *Least-Squares Value Iteration with Upper Confidence Bound* (LSVI-UCB), is given in Algorithm 1.

Algorithm 1 LSVI-UCB

for episode $k = 1, 2, \ldots, K$ **do**

1. (Value estimation) **for** step $h = H, H-1, \ldots, 1$ **do**
   (a) (Gram matrix) $\Lambda_h^k \leftarrow \sum_{\tau\in[k-1]} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + I$
   (b) (Least squares)
$$\begin{aligned} w_h^k &\leftarrow \arg\min_{w\in\mathbb{R}^d} \sum_{\tau\in[k-1]} \left[r_h(x_h^\tau, a_h^\tau) + V_{h+1}^k(x_{h+1}^\tau) - \langle w, \phi(x_h^\tau, a_h^\tau)\rangle\right]^2 + \|w\|^2 \\ &= (\Lambda_h^k)^{-1} \sum_{\tau\in[k-1]} \phi(x_h^\tau, a_h^\tau)\cdot\left[r_h(x_h^\tau, a_h^\tau) + V_{h+1}^k(x_{h+1}^\tau))\right]. \end{aligned}$$
   (c) (Q estimate with UCB) $Q_h^k(\cdot,\cdot) = \left\langle w_h^k, \phi(\cdot,\cdot)\right\rangle + \beta\sqrt{\phi(\cdot,\cdot)^\top(\Lambda_h^k)^{-1}\phi(\cdot,\cdot)}$
   (d) (From Q to value function) $V_h^k(\cdot) = \max_a Q_h^k(\cdot, a)$.
2. Receive initial state x_1^k
3. (Policy execution) **for** step $h = 1, 2, \ldots, H$ **do**
   Take action $a_h^k \leftarrow \arg\max_a Q_h^k(x_h^k, a)$; observe reward $r_h^k = r_h(x_h^k, a_h^k)$ and next state x_{h+1}^k.

Below we discuss and provide intuition for the steps in Algorithm 1.

[1] *Reading:* [Jin et al., 2019]
[2] This normalization ensures consistency when reducing to tabular case.

3.1 Least squares estimation (Step 1(a)–(b))

Recall that $Q_h^*(x,a) = \langle \phi(x,a), w_h^* \rangle$. Our first goal is to estimate the unknown w_h^* associated with the optimal Q_h^*. Under the linear assumption, Lemma 1 guarantees that

$$\langle \phi(x,a), w_h^* \rangle = r_h(x,a) + (\mathbb{P}_h V_{h+1}^*)(x,a).$$

The next-step value function V_{h+1}^* on the RHS is unknown, so we may replace it by V_{h+1}^k, which is our current estimate of the next-step value function. The conditional distribution $\mathbb{P}_h(x'|x,a)$ is also unknown, but we can estimate it using empirical observation of the next state x' (this is called a one-sample estimate). Combining, we see that w_h^* satisfies the following approximate relationship:

$$\langle \phi(x,a), w_h^* \rangle \approx r_h(x,a) + V_{h+1}^k(x').$$

Therefore, we can estimate w_h^* by finding a vector w_h^k that minimizes the difference between the LHS and RHS of the above, over the data from episode $1, \cdots, k-1$. That is,

$$w_h^k \leftarrow \arg\min_{w \in \mathbb{R}^d} \sum_{\tau \in [k-1]} \left[r_h(x_h^\tau, a_h^\tau) + V_{h+1}^k(x_{h+1}^\tau) - \langle w, \phi(x_h^\tau, a_h^\tau) \rangle \right]^2 + \|w\|^2 .$$

The regularization term $\|w\|^2$ ensures uniqueness of the solution w_h^k. The optimal solution of w_h^k can be written in a closed-form as shown in step 1(b) in Algorithm 1.

3.2 Value estimation and bonus term (Step 2(c))

Given the estimate w_h^k, we can calculate $Q_h^k(\cdot,\cdot)$ and $V_h^k(\cdot)$, which are estimates of the true value and Q functions Q_h^* and V_h^*, in Steps 1(c) and 1(d) respectively.

In step 1(c) above, we add a "bonus" term $\beta\sqrt{\phi(\cdot,\cdot)^\top (\Lambda_h^k)^{-1} \phi(\cdot,\cdot)}$ that accounts for the uncertainty in the least square estimate w_h^k, thereby ensuring that with high probability $Q_h^k(x,a)$ is an upper confidence bound (UCB) of the true Q function $Q_h^*(x,a)$ for all (x,a). This upper bound is larger for state-action pairs (x,a) that are infrequently visited in the past, so it encourages exploration of these pairs in Step 3. This idea, as well as the particular form of the bonus, are a generalization of the UCB algorithm for multi-arm bandit.

Recall that tabular MDP is a special case of the linear MDP. It is instructive to look at the particular form of the "bonus" term in the tabular setting. In this setting, $d = |\mathcal{S}||\mathcal{A}|$ and the feature map $\phi(s,a) = \mathbf{e}_{xa}$, which takes 1 at the xa entry and 0 at the other entries. Consequently, the gram matrix

$$\Lambda_h^k = I + \sum_{\tau}^{k-1} \mathbf{e}_{x_h^\tau a_h^\tau} \mathbf{e}_{x_h^\tau a_h^\tau}^\top \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$$

is a diagonal matrix, where each diagonal entry is

$$\begin{aligned}
\Lambda_h^k(xa, xa) &= 1 + \sum_{\tau=1}^{k-1} \mathbb{1}\{(x_h^\tau, a_h^\tau) = (x,a)\} \\
&= 1 + \#\text{ of visits to } (x,a) \text{ pair in step } h \text{ of episode } 1, \cdots, k-1 \\
&=: 1 + N^{k-1}(x,a).
\end{aligned}$$

Thus, the bonus term in the tabular case takes the form

$$\sqrt{\phi(x,a)^\top (\Lambda_h^k)^{-1} \phi(x,a)} = \sqrt{\frac{1}{1 + N^{k-1}(x,a)}}.$$

This implies that the fewer visits to the state-action pair (x,a), the larger the "bonus" term for that pair, which aligns with our initial intention for the "bonus" term.

In the general linear MDP case, we may have two different state-action pairs with similar features $\phi(x,a) \approx \phi(x',a')$. In this case, they will have similar "bonus" terms.

3.3 Policy Execution (Step 2 & 3)

Given $Q_h^k(\cdot,\cdot), h \in [H]$, we can construct the corresponding (deterministic) greedy policy $\pi^k = (\pi_h^k)_{h\in[H]}$, where $\pi_h^k(x) = \arg\max_a Q_h^k(x,a)$. Starting from the initial state x_1^k in Step 2, we can then execute the policy π^k to play out episode k, as shown in Step 3.

3.4 Computational complexity

Regarding the computational complexity of Algorithm 1, it cannot be implemented if we *naively* follow the steps since when $|\mathcal{S}| = \infty$, we have infinite state-action pairs and it's impossible to calculate $Q_h^k(\cdot,\cdot)$ and $V_h^k(\cdot)$ for each pair in each k and h.

A closer inspection of Algorithm 1 shows that we only need to calculate $Q_h^k(x,\cdot)$, $V_h^k(x)$ for the states x that we actually encounter during the episodes.

4 Regret bound

We establish the following regret bound for LSVI-UCB. Here $T := KH$ is the total number of steps over all episodes.

Theorem 4. *Set $\beta = cdH\sqrt{\iota}$ with $\iota := \log(2dT/p)$. With probability at least $1-p$, we have*

$$\mathrm{Regret}(K) := \sum_{k=1}^{K}\left[V_1^*(x_1^k) - V_1^{\pi^k}(x_1^k)\right] \lesssim \sqrt{d^3H^3T\iota^2} = \sqrt{d^3H^4K\iota^2}.$$

Remark 5. We can compare the above regret bound with some known minimax bounds. For tabular MDP, the minimax regret bound is regret $\gtrsim \sqrt{dH^3K}$. Therefore, the dependence on H in Theorem 4 is off by a factor of $\sqrt{H}$.

For linear bandit problem where $H = 1$, the minimax regret bound is regret $\gtrsim \sqrt{d^2K}$. Therefore, the dependence on d in Theorem 4 is off by a factor of $\sqrt{d}$. We will point out later where this additional d factor comes from in the proof.

4.0.1 Sample complexity bound

From the regret bound in Theorem 4, we can derive a sample complexity bound for finding an ϵ-optimal policy. Algorithm 1 outputs K policies $\pi^1, \pi^2, \cdots, \pi^K$. Among them we can randomly pick a policy: $\hat{\pi} \sim \mathrm{uniform}\{\pi^1,\cdots,\pi^K\}$. For a given $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{P}(V_1^*(x_1) - V_1^{\hat{\pi}}(x_1) \geq \epsilon) &\leq \frac{\mathbb{E}[V_1^*(x_1) - V_1^{\hat{\pi}}(x_1)]}{\epsilon} && \text{(by Markov Inequality)}\\
&= \frac{\frac{1}{K}\sum_{k=1}^{K}[V_1^*(x_1) - V_1^{\hat{\pi}}(x_1)]}{\epsilon}\\
&\leq \frac{\frac{1}{K}\sqrt{d^3H^4K}}{\epsilon} && \text{(ignore } \iota \text{ in Theorem 4)}\\
&= \sqrt{\frac{d^3H^4}{K\epsilon}}
\end{aligned}$$

It follows that when $K \geq \frac{d^3H^4}{0.1^2\epsilon}$, we have $\mathbb{P}(V_1^*(x_1) - V_1^{\hat{\pi}}(x_1) \geq \epsilon) \leq 0.1$. This means that with probability ≥ 0.9, $\hat{\pi}$ is an ϵ-optimal policy.

Remark 6. When specialized to the tabular setting where $d = |\mathcal{S}||\mathcal{A}|$, the above sample complexity bound for LSVI-UCB becomes $K \gtrsim \frac{|\mathcal{S}|^3|\mathcal{A}|^3H^4}{\epsilon^2}$. One may compare this bound with the minimax sample complexity bound we get last week, which reads $K \gtrsim \frac{1}{(1-\gamma)^3}\cdot\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon^2} \approx H^3\cdot\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon^2}$, where we treats the effective horizon $\frac{1}{1-\gamma}$ as the horizon. We see that the above sample complexity bound for LSVI-UCB is sub-optimal by a factor of $H\cdot|\mathcal{S}|^2|\mathcal{A}|^2$ not an optimal bound. The additional H factor can be removied by changing the current "Hoeffding bonus" term to a "Bernstein Bonus" term. It is not clear yet whether the difference in $|\mathcal{S}||\mathcal{A}|$ can be removed.

5 Proof of Theorem 4

The proof proceeds in 5 steps.

1. Concentration
2. Least-squares estimation error
3. UCB property
4. Regret decomposition
5. Final regret bound

Today we will cover Step 1.

Define the shorthand $\phi_h^\tau := \phi(x_h^\tau, a_h^\tau)$.

5.1 Concentration

We present a concentration result, which is the crucial step of the proof. For a given positive definite matrix A, define the weighted norm $\|u\|_A := \sqrt{u^\top A u}$.

Lemma 7 (Concentration of empirical measure)**.** *For each p, the following event $\mathfrak{E}$ holds with probability at least $1 - p/2$:*

$$\left\| \sum_{\tau \in [k-1]} \phi_h^\tau \Big[V_{h+1}^k(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^k)(x_h^\tau, a_h^\tau) \Big] \right\|_{(\Lambda_h^k)^{-1}} \lesssim dH\sqrt{\log(dT/p)}, \qquad \forall k, h.$$

Roughly speaking, this lemma says that the empirical sum $\sum_\tau \phi_h^\tau \cdot V(x_{h+1}^\tau)$ approximates the true expectation $\sum_\tau \phi_h^\tau \cdot (\mathbb{P}_h V)(x_h^\tau, a_h^\tau)$. The approximation error is measured in the norm $\|\cdot\|_{(\Lambda_h^k)^{-1}}$ weighted by the Gram matrix $\Lambda_h^k := I + \sum_{\tau\in[k-1]} \phi_h^\tau (\phi_h^\tau)^\top$, where we recall that $\phi_h^\tau = \phi(x_h^\tau, a_h^\tau)$ are feature vectors of the previous visited state-action pairs (x_h^τ, a_h^τ). Therefore, we have better approximation in the directions that are better covered by the previous data. Here we crucially exploit the linear structure: we care about coverage w.r.t. the feature space rather than w.r.t. individual state-action pairs.

Proof Fix k and h. For each $\tau \in [k]$, define the sigma-algebra $\mathcal{F}_{\tau-1} = \sigma(x_{1:H}^1, \ldots x_{1:H}^{\tau-1}, x_1^\tau, \ldots, x_h^\tau)$, which includes everything up to step h of episode τ. Note that $\phi_h^\tau, x_h^\tau \in \mathcal{F}_{\tau-1}$ and $x_{h+1}^\tau \in \mathcal{F}_\tau$.

Consider V_{h+1}^k as fixed first. Note that $V_{h+1}^k(x_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(x_h^\tau) \mid \mathcal{F}_{\tau-1}$ is zero-mean and H-bounded. Applying the concentration inequality for self-normalized processes (Lemma 10), we obtain that with probability at least $1 - \delta$:

$$\left\| \sum_{\tau \in [k-1]} \phi_h^\tau \Big[V_{h+1}^k(x_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(x_h^\tau, a_h^\tau) \Big] \right\|_{(\Lambda_h^k)^{-1}} \lesssim H\sqrt{\log \frac{(k+1)^{d/2}}{\delta}}. \tag{1}$$

Note that the log factor on the RHS comes from the bound $\det \Lambda_h^k \le \left(\left\| \Lambda_h^k \right\|_{\mathrm{op}} \right)^d \le (k+1)^d$.

In reality, V_{h+1}^k is random. By construction and Lemma 9, V_{h+1}^k must lie in the set

$$\begin{aligned} \mathcal{V} := \Big\{ V : V(\cdot) = \max_a \Big[\phi(\cdot, a)^\top w + \beta \sqrt{\phi(\cdot, a) \Lambda^{-1} \phi(\cdot, a)} \Big], \\ w \in \mathbb{R}^d \text{ with } \|w\| \le H\sqrt{dk}, \Lambda \in \mathbb{R}^{d\times d} \text{ with } \lambda_{\min}(\Lambda) \ge 1. \Big\} \end{aligned}$$

The ϵ-covering number of $\mathcal{V}$ is $N \approx \left(\frac{H\sqrt{dk}}{\epsilon}\right)^d \left(\frac{\beta\sqrt{d}}{\epsilon}\right)^{d^2}$, since we need to cover the sets $\left\{ w \in \mathbb{R}^d : \|w\| \le H\sqrt{dk} \right\}$ and $\left\{ A \in \mathbb{R}^{d\times d} : A = \Lambda^{-1}, \lambda_{\max}(A) = \lambda_{\min}(\Lambda)^{-1} \le 1 \right\}$.

Applying (1) with $\delta = \frac{p/2}{N}$ and running an ϵ-net argument to all possible V_{h+1}^k in $\mathcal{V}$, we obtain the desired inequality. $\square$

Note that the d factor on the RHS of the lemma statement comes from $\sqrt{\log N} \approx d$.

Appendices

A Technical lemmas

We begin with a simple upper bound on the Gram matrix.

Lemma 8 (Simple upper bound). *If $\Lambda_t = \lambda I + \sum_{i\in[t]} \phi_i \phi_i^\top \in \mathbb{R}^d$ and $\lambda > 0$, then*

$$\sum_{i\in[t]} \phi_i^\top \Lambda_t^{-1} \phi_i \leq d.$$

Proof If $\lambda = 0$, then it is easy to see that $\sum_{i\in[t]} \phi_i^\top \Lambda_t^{-1} \phi_i = \mathrm{tr}(I_d) = d$. The regularization $\lambda > 0$ only makes the LHS smaller. $\square$

The next lemma ensures boundedness of the linear weights.

Lemma 9 (Weights are bounded). *(i) For each policy π and its Q function $Q_h^\pi(x,a) = \langle \phi(x,a), w_h^\pi \rangle$, we have $\|w_h^\pi\| \leq 2H\sqrt{d}, \forall h$. (ii) The weights $\{w_h^k\}$ in the LSVI-UCB algorithm satisfies $\left\|w_h^k\right\| \leq 2H\sqrt{dk}, \quad \forall k, h$.*

Proof Part (i) follows from Assumption 1 on linearity and boundedness.Part (ii) holds since the Gram matrix Λ_h^k has minimum eigenvalue ≥ 1 and satisfies the bound in Lemma 8. $\square$

Lemma 10 (Concentration for self-normalized processes [Abbasi-Yadkori et al., 2011, Theorem 1]). *Suppose $(\epsilon_s)_{s=1,2,\ldots}$ is a scalar stochastic process adapted to the filtration $(\mathcal{F}_s)$, and $\epsilon_s | \mathcal{F}_{s-1}$ is zero mean and σ-sub-Gaussian. Let $(\phi_s)_{s=1,2,\ldots}$ be an $\mathbb{R}^d$-valued stochastic process with $\phi_s \in \mathcal{F}_{s-1}$. Let $\Lambda_t = I + \sum_{s=1}^t \phi_s \phi_s^\top \in \mathbb{R}^{d\times d}$. Then with probability at least $1-\delta$, we have*

$$\left\| \sum_{s=1}^t \phi_s \epsilon_s \right\|_{\Lambda_t^{-1}}^2 \leq 2\sigma^2 \log\left[\frac{\det(\Lambda_t)^{1/2}}{\delta}\right], \qquad \forall t \geq 0.$$

References

[Abbasi-Yadkori et al., 2011] Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, pages 2312–2320.

[Jin et al., 2019] Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2019). Provably efficient reinforcement learning with linear function approximation.