Lecture 9: Exploration with linear function approximation

Akshay Krishnamurthy
akshay@cs.umass.edu

March 29, 2022

In the last lecture we discussed the challenges of exploration and the UCB-VI algorithm, which uses optimistic dynamic programming with confidence bonuses to achieve low regret in the tabular markov decision process. This algorithm addresses two of the three core capabilities for reinforcement learning. Today, we'll combine UCB-VI with techniques from linear bandits to address all three capabilities.

Recall the usual definitions: we have an MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \mu)$ with horizon H and we consider the learning/exploration setting where the agent interacts with the MDP for T episodes. In the t^{th} episode we generate a trajectory $\tau_t := (s_0^t, a_0^t, r_0^t, s_1^t, a_1^t, r_1^t, \ldots, s_{H-1}^t, a_{H-1}^t, r_{H-1}^t)$ where actions are chosen by the agent and we measure performance via regret:

$$\mathrm{Reg}(T) := TJ(\pi^\star) - \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1} r_h^t\right].$$

In exactly this setup we saw that UCB-VI can obtain a regret bound scaling as $\mathrm{poly}(S, A, H)\sqrt{T}$. Today, the goal will be to avoid dependence on S and A through the use of linear function approximation.

1 The linear MDP

To enable linear function approximation, we assume access to a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ that we will use to approximate quantities of interest. We will not precisely specify any assumptions at this point but drawing inspiration from linear bandits, it seems natural to assume that $Q^\star$ is linearly realizable by the features. Indeed, this specializes to the stochastic linear bandits problem if $H = 1$.

Drawing inspiration from UCB-VI, it may be natural to use the features in the optimistic dynamic programming procedure we had previously. Recall that for UCB-VI we performed the iteration:

$$Q_{H-1}^{t-1}(s,a) = R(s,a), \qquad Q_h^{t-1}(s,a) = \min\{H, R(s,a) + b^{t-1}(s,a) + \sum_{s'} P^{t-1}(s' \mid s,a) \max_{a'} Q_{h+1}^{t-1}(s',a')\}.$$

Here, while we did estimate the transition operator P^{t-1} from samples, we mostly care about the optimistic Q function itself. So it may be natural to replace the "model-based" nature with a dynamic programming procedure that directly works with Q functions. Specifically, at episode t we could use all of the previous data to perform the dynamic programming scheme. Starting with $V_H^t \equiv 0$ we compute

$$\forall h \le H-1 : \theta_h^t \leftarrow \operatorname*{argmin}_\theta \sum_{i=1}^{t-1} \left(\left\langle \phi(s_h^i, a_h^i), \theta \right\rangle - r_h^i - V_{h+1}^t(s_{h+1}^i)\right)^2,$$
$$Q_h^t : (s,a) \mapsto \min\{H, \left\langle \phi(s,a), \theta_h^t \right\rangle + b_h^{t-1}(s,a)\}, \qquad V_h^t : s \mapsto \max_a Q_h^t(s,a)$$

where b_h^{t-1} is a bonus function that encourages optimism. Then we can simply deploy the greedy policy with respect to Q^t. This is very similar to the UCB-VI procedure, except we use linear regression to approximate the Bellman backups, instead of a more direct tabular approach.

Today we will analyze essentially this algorithm, but we will need to make fairly strong assumptions on the MDP to prove that it succeeds. Thinking about this algorithm, we immediately encounter an obstacle in that the

regression problems we are solving may not be well-specified/realizable. Indeed this will be true even if we assume that $Q^\star$ is realizable by our features, since our regression targets are not unbiased estimates of $Q^\star$! There are two sources of error here, one is simply that $Q_h^\star(s,a) \neq \langle \phi(s,a), \theta_h^t \rangle$ due to statistical fluctuations, and the other is that we are adding some bonus to ensure optimism.

While one can obtain linear regression guarantees without realizability, they are not very useful for establishing optimism. For example, one guarantee you can obtain takes the form $L(\hat{\theta}) - \min_\theta L(\theta) \leq \varepsilon_{\text{stat}}$, where L is the population square loss and $\varepsilon_{\text{stat}}$ goes to zero with the amount of data. However, under misspecification, the best parameter may not be the one that realizes the conditional mean of the target (which in this case is $\mathcal{T}V_{h+1}^t$), so we may have some irreducible error. It is not clear how to establish optimism in this case and later on we will see that exploration with linear function approximation is not possible without much stronger assumptions.

On the other hand, if the regression problem were well-specified, we already learned how to guarantee optimism when we studied stochastic linear bandits. From a technical perspective, this motivates the following *linear MDP* assumption, under which we can establish realizability of the regression problems.

Definition 1 (Linear/Low rank MDP)**.** *An MDP $M = (\mathcal{S}, \mathcal{A}, P, R)$ is a* linear MDP *if there is a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, unknown vector $w^\star : \mathbb{R}^d$, and unknown signed measures, $\mu : \mathcal{S} \to \mathbb{R}^d$ such that:*

$$R(s,a) = \langle \phi(s,a), w^\star \rangle, \qquad P(s' \mid s,a) = \langle \phi(s,a), \mu(s') \rangle.$$

The standard regularity/normalization assumptions are: $\sup_{s,a} \|\phi(s,a)\|_2 \leq 1$, $\sup_{f:\|f\|_\infty \leq 1} \|\int \mu(s) f(s)\|_2 \leq \sqrt{d}$, $\|w^\star\|_2 \leq W$, and rewards are in $[0,1]$.

We can also think of the μ mapping as a matrix $\mu \in \mathbb{R}^{S \times d}$ such that $P(\cdot \mid s,a) = \mu\phi(s,a)$. Then the normalization condition is that $\sup_{v:\|v\|_\infty \leq 1} \|v^\top \mu\|_2 \leq \sqrt{d}$. Note that any MDP can be written as a linear MDP with $d \leq SA$ and so linear MDP methods generalize tabular ones.

The linear MDP asserts that the reward function is linearly realizable in the features, which we also assumed in stochastic linear bandits, so this is fairly natural. Here we also assume that the transition operator satisfies some linearity property. Observe that this is equivalent to writing $P = \mu\Phi$ where $\mu \in \mathbb{R}^{S \times d}$ and $\Phi \in \mathbb{R}^{d \times SA}$, which reveals the low rank structure in the transition operator. This assumption may seem more unnatural, but it admits the following favorable property, which addresses the realizability issue above.

Proposition 2. *For any function $f : \mathcal{S} \to [0,H]$, there exists $\theta_f \in \mathbb{R}^d$ with $\|\theta_f\|_2 \leq H\sqrt{d}$ such that*

$$\forall s,a : \mathbb{E}_{s' \sim P(s,a)}[f(s')] = \langle \phi(s,a), \theta_f \rangle$$

From this proposition, we can immediately see that the regression problems in the optimistic dynamic programming update are now well-specified, which will make it much easier for us to establish optimism.

Proof. Fix (s,a) and observe

$$\mathbb{E}_{s' \sim P(s,a)}[f(s')] = \int_{s'} P(s' \mid s,a) f(s') = \int_{s'} \langle \phi(s,a), \mu(s') \rangle f(s') = \left\langle \phi(s,a), \int_{s'} \mu(s') f(s') \right\rangle.$$

The latter vector, which we call θ_f, has ℓ_2 norm at most $H\sqrt{d}$ by our regularity assumptions. □

2 LSVI-UCB

Now that we have admittedly strong assumptions, we can introduce and analyze the LSVI-UCB algorithm. The algorithm is almost what we presented above, except we use ridge regularization in the dynamic programming and we must also specify the bonuses. At the beginning of episode t we compute the Q^t function via:

$$\forall h \leq H-1 : \theta_h^t \leftarrow \operatorname*{argmin}_\theta \sum_{i=1}^{t-1} \left(\langle \phi(s_h^i, a_h^i), \theta \rangle - r_h^i - V_{h+1}^t(s_{h+1}^i) \right)^2 + \lambda \|\theta\|_2^2,$$
$$Q_h^t : (s,a) \mapsto \min\{H, \langle \phi(s,a), \theta_h^t \rangle + b_h^{t-1}(s,a)\}, \qquad V_h^t : s \mapsto \max_a Q_h^t(s,a),$$

starting with $V_H^t \equiv 0$ and setting the bonus as $b_h^{t-1}(s,a) = \beta\|\phi(s,a)\|_{\Lambda_{h,t-1}^{-1}}$ where $\Lambda_{h,t} = \sum_{i=1}^{t-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$ and $\beta = \tilde{O}(Hd)$ will be specified later.

For intuition, if we take $H = 1$ this is *exactly* the LinUCB algorithm we studied earlier in the course. In particular, we used ridge regularization and set the optimism bonus in terms of the inverse covariance of the observed features. So these components are identical. The only difference is that we now also perform dynamic programming, which is how we can scale to longer horizon problems.

Analyzing this algorithm essentially combines the analysis of of UCB-VI and LinUCB. In particular we have three steps: (a) optimistic regret decomposition (which is exactly the same as for UCB-VI), (b) establishing optimism (which is close to LinUCB), and (c) bounding the confidence sum (which is exactly the same as LinUCB). Let us sketch out these three parts.

Optimistic regret decomposition. Last lecture we proved that if we have $\{Q_h^t, \mathrm{conf}_h^t\}_{h,t}$ satisfying

$$\forall s, a, h, t : Q_h^\star(s,a) \le Q_h^t(s,a) \le (\mathcal{T}Q_{h+1}^t)(s,a) + \mathrm{conf}_h^t(s,a),$$

and we deploy the greedy policy with respect to Q_h^t, then with probability at least $1-\delta$ we have the regret bound

$$\mathrm{Reg}(T) \le \sum_t \sum_h \mathrm{conf}_h^t(s_h^t, a_h^t) + O(H\sqrt{T\log(1/\delta)}).$$

Establishing optimism. Similar to last time we establish optimism inductively. Consider episode t and time step h and assume that $V_{h+1}^t(s') \ge V_{h+1}^\star(s')$ for all s'. Then observe that our regression targets satisfy

$$\mathbb{E}[r + V_{h+1}^t(s') \mid s, a] = \langle \phi(s,a), w^\star \rangle + \left\langle \phi(s,a), \bar{\theta}_h^t \right\rangle = \left\langle \phi(s,a), \tilde{\theta}_h^t \right\rangle$$

where $\bar{\theta}_h^t$ is the parameter vector for V_{h+1}^t promised by Proposition 2 and $\tilde{\theta}_h^t = w^\star + \bar{\theta}_h^t$. Thus, we have a well-specified regression problem and can deduce

$$\|\theta_h^t - \tilde{\theta}_h^t\|_{\Lambda_{h,t-1}}^2 \le \tilde{O}(H^2 d^2) =: \beta. \tag{1}$$

In stochastic linear bandits, we stated (but did not prove) a similar inequality, with a right hand side of $\tilde{O}(\sqrt{d})$. Instead here we have a right hand side of $\tilde{O}(H\sqrt{d})$. The factor of H arises from the range of the value functions which influence the scale of the noise.

The additional factor of d arises for a more subtle reason. While the regression problem is well specified, actually the optimal predictor $\tilde{\theta}_h^t$ is *random* since it is determined by V_{h+1}^t which itself depends on all of the past episodes. This couples all of the terms in the square loss regression problem together so that they are no longer independent which is important in the concentration argument. We resolve this issue by doing a *uniform convergence* argument, essentially considering all possible choices for V_{h+1}^t and taking a large union bound. This incurs an extra $\sqrt{d}$ factor because V_{h+1}^t is determined by coefficients θ_{h+1}^t and the second moment matrix $\Lambda_{h+1,t-1} \in \mathbb{R}^{d\times d}$ so we have to take a union bound over roughly $\exp(d^2)$ choices for V_{h+1}^t.

Taking the bound in (1) as true, we can establish the conditions of the optimistic regret decomposition:

$$\begin{aligned} Q_h^\star(s,a) &= \langle \phi(s,a), w^\star \rangle + (\mathcal{T}V_{h+1}^\star)(s,a) \le \langle \phi(s,a), w^\star \rangle + \left\langle \phi(s,a), \bar{\theta}_h^t \right\rangle = \left\langle \phi(s,a), \tilde{\theta}_h^t \right\rangle \\ &= \left\langle \phi(s,a), \theta_h^t \right\rangle + \left\langle \phi(s,a), \tilde{\theta}_h^t - \theta_h^t \right\rangle \le \left\langle \phi(s,a), \theta_h^t \right\rangle + \|\phi(s,a)\|_{\Lambda_{h,t-1}^{-1}} \cdot \beta = Q_h^t(s,a). \end{aligned}$$

The first inequality uses our inductive hypothesis that V_{h+1}^t is optimistic, since $\left\langle \phi(s,a), \tilde{\theta}_h^t \right\rangle = (\mathcal{T}V_{h+1}^t)(s,a) \ge (\mathcal{T}V_{h+1}^\star)(s,a)$. The second uses Cauchy-Schwarz and the regression guarantee. The upper inequality is similar:

$$Q_h^t(s,a) \le \left\langle \phi(s,a), \theta_h^t \right\rangle + \beta\|\phi(s,a)\|_{\Lambda_{h,t-1}^{-1}} \le \left\langle \phi(s,a), \tilde{\theta}_h^t \right\rangle + 2\beta\|\phi(s,a)\|_{\Lambda_{h,t-1}^{-1}} = (\mathcal{T}V_{h+1}^t)(s,a) + 2\beta\|\phi(s,a)\|_{\Lambda_{h,t-1}^{-1}},$$

which shows that we can set $\mathrm{conf}_h^t(s,a) = 2\beta\|\phi(s,a)\|_{\Lambda_{h,t-1}^{-1}}$.

Potential function. Based on the above two steps, we can bound the regret by

$$\mathrm{Reg}(T) \le 2\beta \sum_t \sum_h \max(H, \|\phi(s_h^t, a_h^t)\|_{\Lambda_{h,t-1}^{-1}}) + \tilde{O}(H\sqrt{T}).$$

Here we can apply the elliptical potential lemma to bound the dominant term as $\tilde{O}(\beta \cdot H\sqrt{dT})$. This leads to the following theorem.

Theorem 3. *Assuming M satisfies the linear MDP property, then with probability at least $1-\delta$, LSVI-UCB has* $\mathrm{Reg}(T) \le \tilde{O}(H^2\sqrt{d^3T})$.

3 Weaker assumptions and linear Bellman completeness

While the linear MDP subsumes tabular MDPs and admits sample efficient reinforcement learning, the assumption is quite strong, so it is reasonable to ask if weaker assumptions remain tractable. An immediate observation is that LSVI-UCB only ever creates regression problems with targets from a specific class of functions, so only these functions need to have linear Bellman backups. In more detail, if we define $\mathcal{G} = \{s \mapsto \max_a \min\{H, \langle\phi(s,a),\theta\rangle + \beta\|\phi(s,a)\|_{M^{-1}}\} : \theta \in \mathbb{R}^d, M \succeq 0, \beta \ge 0\}$, then we only need that $\mathcal{T}g$ is a linear function in $\phi(s,a)$ for $g \in \mathcal{G}$. This is a strictly weaker than the linear MDP, which requires that *all* functions admit linear Bellman backups.

Once you move away from explicit assumptions on the transition model, you can also start to move away from linear function approximation. The main other component required by LSVI-UCB is that we can build pointwise confidence intervals that shrink rapidly. There are some nonlinear function classes for which this is also true, and this property is typically captured via the notion of Eluder dimension, which you may see in the literature. One simple nonlinear example is a generalized linear model where we approximate $Q(s,a) = \sigma(\langle\phi(s,a),\theta\rangle)$ for some *link function* σ that is continuous and monotone. For example, we could take $\sigma(x) = 1/(1+\exp(x))$ to be the sigmoid function. For generalized linear functions, essentially the same algorithm and analysis apply. (There are some tricks required for working with more general Eluder classes.)

However, even the weaker property that $\mathcal{TG}$ is linearly realizable for the class $\mathcal{G}$ defined above is somewhat strong and precludes some models of interest. For example, it precludes the linear quadratic regulator, which is the canonical model in control theory. Instead, the most natural assumption is what is known as "Bellman completeness" or "closure of the bellman operator":

Assumption 4 (Bellman completeness)**.** *A function class $\mathcal{F} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and an MDP M satisfy* bellman completeness *if $\forall f \in \mathcal{F}$ we have $\mathcal{T}f \in \mathcal{F}$.*

If we take $\mathcal{F}$ to be the class of linear functions in the features ϕ then we are asking that all linear functions in ϕ admit linear-in-ϕ Bellman backups. This is a weaker assumption than the linear MDP and is referred to as the "linear Bellman completeness" setting. Note that LSVI-UCB does not seem to work under these weaker assumptions, since it creates misspecified regression problems, for which it is difficult to establish optimism. Can we develop efficient algorithms for this setting?

The answer turns out to be yes on the statistical side, but it currently remains open to obtain a computationally efficient algorithm. The algorithm also highlights a distinction between local and global optimism, which is the main technical novelty. In LSVI-UCB we ensured that Q_h^t is pointwise optimistic which lead to the optimistic regret decomposition. It turns out that this is not necessary and it suffices to have $(Q_0^t, \dots, Q_{H-1}^t)$ satisfy a more global optimism property, simply that $\mathbb{E}_{s_0\sim\mu}[\max_a Q_0^t(s_0,a) - \max_a Q_0^\star(s_0,a)] \ge 0$. This leads to a different regret decomposition:

Lemma 5 (Global optimistic regret decomposition)**.** *Suppose we have a Q function $(Q_0, \dots, Q_{H-1})$ such that $\mathbb{E}_{s_0}\max_a Q_0(s_0,a) \ge \mathbb{E}_{s_0}\max_a Q_0^\star(s_0,a)$ and we set π to be the greedy policy with respect to Q. Then*

$$J(\pi^\star) - J(\pi) \le \sum_{h=0}^{H-1} \mathbb{E}_{(s,a)\sim d_h^\pi}[Q_h(s,a) - (\mathcal{T}Q_{h+1})(s,a)]$$

Observe that the requirement on the Q-function is significantly milder than the pointwise optimistic property.

Proof.

$$\begin{aligned}
J(\pi^\star) - J(\pi) &= \mathbb{E}_{s_0}\left[\max_a Q_0^\star(s_0,a) - Q_0^\pi(s_0,\pi(s_0))\right] \\
&\le \mathbb{E}_{s_0}\left[\max_a Q_0(s_0,a) - Q_0^\pi(s_0,\pi(s_0))\right] \\
&= \mathbb{E}_{(s_0,a_0)\sim d_0^\pi}\left[Q_0(s_0,a_0) - Q_0^\pi(s_0,a_0)\right] \\
&= \mathbb{E}_{(s_0,a_0)\sim d_0^\pi}\left[Q_0(s_0,a_0) - (\mathcal{T}Q_1)(s_0,a_0)\right] + \mathbb{E}_{s_1\sim d_1^\pi}\left[\max_a Q_1(s_1,a) - Q_1^\pi(s_1,\pi(s_1))\right].
\end{aligned}$$

The first inequality uses the global optimism property and from then on we use the fact that π is greedy with respect to Q and the definition of the Bellman operator. Note that the first term appears on the RHS of the lemma statement, while the second term is the same as the expression on the second line one time step in the future, so we can unroll this argument. (Observe that this proof is identical to the proof for the previous regret decomposition although the statement is different.) □

The main challenge now is to ensure global optimism assuming only bellman completeness. This is where the computational issues arise. The problem is that if we do regression, due to statistical errors, we will not obtain $Q^\star$ exactly. If we then use our estimate as the regression target at the previous time the solution may be very different from $Q^\star$ and we cannot add bonuses to correct for this discrepancy.

We resolve this by maintaining confidence balls for plausible parameters. Let us define

$$R_h^{t-1}(\theta,\tilde{\theta}) := \sum_{i=1}^{t-1}\left(\left\langle\phi(s_h^i,a_h^i),\theta\right\rangle - r_h^i - \max_a\left\langle\phi(s_{h+1}^i,a),\tilde{\theta}\right\rangle\right)^2 + \lambda\|\theta\|_2^2,$$

which is the regression error at time step h if we use $\tilde{\theta}$ to compute the regression targets. Then we define a feasible set of parameter tuples:

$$\textsc{Ball}^t := \left\{(\theta_0,\ldots,\theta_H) : \theta_H = 0, \forall h : R_h^{t-1}(\theta_h,\theta_{h+1}) \le \min_\theta R_h^{t-1}(\theta,\theta_{h+1}) + \beta^2\right\}$$

Thus, we are asking that each parameter θ_h is a near-optimizer for the regression problem defined by the subsequent parameter θ_{h+1}. This is a form of "temporal consistency" in our parameters. If we set β to account for statistical errors, we can see that $(\theta_0^\star,\ldots,\theta_{H-1}^\star,\theta_H^\star) \in \textsc{Ball}^t$, since $\theta_{H-1}^\star$ is the population minimizer for regressing onto the rewards, and since $\theta_h^\star$ is the population minimizer when we regress onto $\theta_{h+1}^\star$. Note also that we are only creating regression problems where the target is a linear function of the features, so under linear Bellman completeness, these are all well-specified.

So in this way, we can achieve global optimism by choosing:

$$\vec{\theta}^t \leftarrow \operatorname*{argmax}_{(\theta_0,\ldots,\theta_H)\in\textsc{Ball}^t} \mathbb{E}_{s_0}\max_a\langle\phi(s_0,a),\theta_0\rangle \tag{2}$$

For the regret analysis, first let us consider some sequence $(\theta_0,\ldots,\theta_H) \in \textsc{Ball}^t$. Let us define $\tilde{\theta}_h$ so that $(\mathcal{T}\theta_{h+1})(s,a) = \left\langle\phi(s,a),\tilde{\theta}_h\right\rangle$. Then since $\tilde{\theta}_h$ is the population minimizer for the regression problem, from the confidence ball property we can derive that $\|\theta_h - \tilde{\theta}_h\|_{\Lambda_{h,t-1}} \le O(\beta)$. This means that the regret is bounded as

$$\begin{aligned}
\mathrm{Reg}(T) &\le \sum_t\sum_h Q_h^t(s_h^t,a_h^t) - (\mathcal{T}Q_{h+1}^t)(s_h^t,a_h^t) + \tilde{O}(H\sqrt{T}) \\
&= \sum_t\sum_h\left\langle\phi(s_h^t,a_h^t),\theta_h^t - \tilde{\theta}_h^t\right\rangle + \tilde{O}(H\sqrt{T}) \\
&\le \sum_t\sum_h\|\phi(s_h^t,a_h^t)\|_{\Lambda_{h,t-1}^{-1}}\cdot\beta + \tilde{O}(H\sqrt{T}) \le \tilde{O}(H\beta\sqrt{dT})
\end{aligned}$$

One upside of this algorithm is that we can set $\beta = \tilde{O}(H\sqrt{d})$ which save a factor of $\sqrt{d}$ when compared with LSVI-UCB. This arises in the uniform convergence argument because, while we do consider a large class of possible regression problems, the set is much smaller since we do not have to account for the bonus. The downside is that it is not clear how to solve the optimization problem in (2), so this algorithm should be viewed as confirmation of statistical tractability even with these weaker assumptions. However the algorithm is not computationally tractable.