Lecture 15

Instructor: Haipeng Luo

1 Stochastic Linear Bandit

In this lecture we introduce yet another classic stochastic bandit model, called *stochastic linear bandit*, and discuss how to use the same principle of "optimism in face of uncertainty" to solve it. There is also a huge literature on this topic and the following discussions follow mostly [Abbasi-Yadkori et al., 2011].

The learning protocol is as follows: for each round $t = 1, \ldots, T$,

1. A set of actions $A_t \subset \mathbb{R}^d$ is revealed to the learner;
2. the learner picks an action $a_t \in A_t$ and observe its loss $c_t = \langle a_t, \theta^\star \rangle + \epsilon_t$ where $\theta^\star \in \mathbb{R}^d$ is an unknown parameter and $\epsilon_t \sim \mathcal{N}(0, 1)$ is independent standard Gaussian noise.

Let $a_t^\star = \operatorname{argmin}_{a \in A_t} \langle a, \theta^\star \rangle$ be the optimal action at time t. The pseudo-regret for this problem is defined as

$$\bar{\mathcal{R}}_T = \mathbb{E}\left[\sum_{t=1}^{T} \langle a_t - a_t^\star, \theta^\star \rangle\right]. \tag{1}$$

First of all, the stochastic multi-armed bandit model we discussed last time is clearly a special case with $d = K$, $A_t = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$ (that is, the standard basis of $\mathbb{R}^d$) and $\theta^\star = (\mu_1, \ldots, \mu_d)$ be the vector of loss means for the actions, except that for simplicity we only consider Gaussian noise now.

In general, the stochastic linear bandit model is much more powerful since it allows each action to come with an arbitrary "feature", and moreover the set of available actions can be different at different time. This allows the model to capture real-life problems such as building a personalized news recommendation system [Li et al., 2010]. In this example, each time t corresponds to a visit of some user to the website. The available news articles at that time as well as the user's information are then used to generate a feature vector for each article. Afterwards a linear bandit algorithm somehow selects an action and recommends the corresponding article to the user. The loss is then based on whether the user clicks on the recommended article or not. It is assumed that the expected loss of an action can be perfectly predicted by an unknown linear predictor $\theta^\star$, but generalization to nonlinear models is possible.

Note that because of the changing action sets, it only makes sense to define the pseudo-regret as in Eq. (1) so that it compares the expected loss of the algorithm to the expected loss of the best action *at each time*. This relates to the notion of dynamic regret discussed before. However, while in general sublinear dynamic regret is impossible, due to the stochastic assumption, regret of order $\mathcal{O}(\sqrt{T})$ is in fact achievable here as we will show soon.

Finally without loss of generality, we make two scaling assumptions: $\max_{a \in A_t} \|a\|_2 \leq 1$ for all t and $\|\theta^\star\|_2 \leq 1$.

2 LinUCB

Let's apply the same "optimism in face of uncertainty" principle to come up with an algorithm for this problem. Recall that the first step is to come up with the set of plausible environments that are

consistent with the observed data. The only parameter of the environment here is the linear predictor $\theta^\star$. So the first goal would be to come up with a confidence set Θ_t based on $a_1, c_1, \ldots, a_t, c_t$ so that $\theta^\star \in \Theta_t$ with high probability. With such a confidence set, similar to the UCB algorithm at time $t+1$ we optimistically assume that the loss for action $a \in A_{t+1}$ is

$$\mathrm{LCB}_{t+1}(a) = \min_{\theta \in \Theta_t} \langle a, \theta \rangle ,$$

and finally pick action $a_{t+1} = \operatorname{argmin}_{a \in A_{t+1}} \mathrm{LCB}_{t+1}(a)$.

It remains to come up with the confidence set Θ_t. First we need to figure out what the "center" of this set is. For UCB, the center of the confidence set is simply and naturally the empirical mean of losses. For linear bandit, note that we are observing $c_\tau \approx \langle a_\tau, \theta^\star \rangle$ for $\tau = 1, \ldots, t$. It is thus natural to perform least square regression to obtain an estimate of $\theta^\star$ as the center:

$$\widehat{\theta}_t = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{\tau=1}^{t} \left(\langle a_\tau, \theta \rangle - c_\tau \right)^2 .$$

By direct calculations one can verify that $\widehat{\theta}_t = M_t^{-1} \sum_{\tau=1}^t c_\tau a_\tau$ where $M_t = \sum_{\tau=1}^t a_\tau a_\tau^\top$ is the covariance matrix and is assumed to be invertible for now. Note that this is consistent with UCB: when $A_t = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$, M_t is a diagonal matrix with $M_t(i,i)$ being the number of times action i has been picked, and $\widehat{\theta}_t$ is exactly the vector of empirical means of actions.

By plugging $c_\tau = \langle a_\tau, \theta^\star \rangle + \epsilon_\tau$, we can also rewrite $\widehat{\theta}_t$ as

$$\widehat{\theta}_t = \left(M_t^{-1} \sum_{\tau=1}^{t} (\langle a_\tau, \theta^\star \rangle + \epsilon_\tau) a_\tau \right) = M_t^{-1} M_t \theta^\star + M_t^{-1} \sum_{\tau=1}^{t} \epsilon_\tau a_\tau = \theta^\star + M_t^{-1} Z_t$$

where $Z_t = \sum_{\tau=1}^t \epsilon_\tau a_\tau$. Next we need to figure out what Θ_t should look like around the center $\widehat{\theta}_t$. To get the intuition, we first ignore the fact that a_τ's are random variables and think of them as fixed vectors (all assumptions mentioned so far will be dropped eventually). Then we have that Z_t is a zero-mean d-dimensional Gaussian variable with covariance matrix

$$\mathbb{E}\left[Z_t Z_t^\top \right] = \sum_{\tau_1=1}^{t} \sum_{\tau_2=1}^{t} \mathbb{E}\left[\epsilon_{\tau_1} \epsilon_{\tau_2} \right] a_{\tau_1} a_{\tau_2}^\top = \sum_{\tau=1}^{t} \mathbb{E}\left[\epsilon_\tau^2 \right] a_\tau a_\tau^\top = M_t .$$

Therefore the random variable $M_t^{1/2}(\widehat{\theta}_t - \theta^\star)$ is actually distributed as $\mathcal{N}(0, I)$, the d-dimensional standard Gaussian. The question thus transfers to finding a region $S \in \mathbb{R}^d$ so that $\Pr(X \in S) \geq 1 - \delta$ if $X \sim \mathcal{N}(0, I)$. By standard results (specifically tail bounds of χ_d^2 distribution), S can be chosen as an ℓ_2-ball with squared radius $d + 2\sqrt{d \ln \frac{1}{\delta}} + 2 \ln \frac{1}{\delta}$. In other words,

$$\Pr\left(\left\| M_t^{1/2} (\widehat{\theta}_t - \theta^\star) \right\|_2^2 \leq d + 2\sqrt{d \ln \tfrac{1}{\delta}} + 2 \ln \tfrac{1}{\delta} \right) \geq 1 - \delta,$$

and the confidence set can thus be

$$\Theta_t = \left\{ \theta \in \mathbb{R}^d : \left\| M_t^{1/2} (\theta - \widehat{\theta}_t) \right\|_2^2 \leq d + 2\sqrt{d \ln \tfrac{1}{\delta}} + 2 \ln \tfrac{1}{\delta} \right\}$$

which is in fact an ellipsoid centered at $\widehat{\theta}_t$. Usually $\left\| M^{1/2} v \right\|_2 = \sqrt{v^\top M v}$ is compactly written as $\|v\|_M$, which is indeed a norm when M is positive definite. The set defined by $\|v\|_M \leq 1$ is the standard analytic form of an ellipsoid centered at the origin. The eigenvectors of M define the principal axes of the ellipsoid while the eigenvalues are the reciprocals of the squares of the semi-axes.

In the process of deriving such a ellipsoidal confidence set, we made two assumptions. First, M_t is invertible, which is not true until $a_1, \ldots, a_t$ span $\mathbb{R}^d$. This can be resolved by adding an ℓ_2-regularization to the least square regression

$$\widehat{\theta}_t = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{\tau=1}^{t} \left(\langle a_\tau, \theta \rangle - c_\tau \right)^2 + \lambda \|\theta\|_2^2 = M_t^{-1} \sum_{\tau=1}^{t} c_\tau a_\tau$$

where M_t is redefined as $\lambda I + \sum_{\tau=1}^{t} a_\tau a_\tau^\top$ for some parameter $\lambda > 0$ and is always invertible now. Similar confidence sets can be constructed based on this new M_t. However, what is more difficult to get rid of is the second assumption that a_t's are fixed and not random. Fortunately, with fancier probability tools, this can still be addressed. Specifically, the following lemma was proven in [Abbasi-Yadkori et al., 2011].

Lemma 1 (Confidence Ellipsoid)**.** *Let* $M_t = \lambda I + \sum_{\tau=1}^{t} a_\tau a_\tau^\top$, $\widehat{\theta}_t = M_t^{-1} \sum_{\tau=1}^{t} c_\tau a_\tau$, $\beta_t = \sqrt{\lambda} + \sqrt{2 \ln \frac{1}{\delta} + d \ln\left(1 + \frac{t}{d\lambda}\right)}$*, and*

$$\Theta_t = \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \widehat{\theta}_t \right\|_{M_t} \leq \beta_t \right\}.$$

Then no matter how a_t*'s are chosen, with probability* $1 - \delta$*,* $\theta^\star \in \Theta_t$ *holds for all* t*.*

Finally, having this confidence set Θ_t, we can further simplify the algorithm by noting

$$\begin{aligned}
\mathrm{LCB}_{t+1}(a) &= \min_{\theta \in \Theta_t} \langle a, \theta \rangle = \min_{\|\theta - \widehat{\theta}_t\|_{M_t} \leq \beta_t} \langle a, \theta \rangle \\
&= \min_{\|\theta'\|_2 \leq \beta_t} \left\langle a, M_t^{-\frac{1}{2}} \theta' + \widehat{\theta}_t \right\rangle \qquad \text{(by changing variable } \theta' = M_t^{\frac{1}{2}}(\theta - \widehat{\theta}_t)\text{)} \\
&= \left\langle a, \widehat{\theta}_t \right\rangle + \min_{\|\theta'\|_2 \leq \beta_t} \left\langle M_t^{-\frac{1}{2}} a, \theta' \right\rangle \\
&= \left\langle a, \widehat{\theta}_t \right\rangle - \beta_t \|a\|_{M_t^{-1}}
\end{aligned}$$

and thus

$$a_{t+1} = \operatorname*{argmin}_{a \in A_{t+1}} \mathrm{LCB}_{t+1}(a) = \operatorname*{argmin}_{a \in A_{t+1}} \left(\left\langle a, \widehat{\theta}_t \right\rangle - \beta_t \|a\|_{M_t^{-1}} \right).$$

This algorithm is called by many names, such as LinUCB or OFUL (Optimism in Face of Uncertainty for Linear bandit). Very similar to UCB, the term $\left\langle a, \widehat{\theta}_t \right\rangle$ drives exploitation while the term $-\beta_t \|a\|_{M_t^{-1}}$ drives exploration of unobserved directions. Indeed, when $A_t = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$, one can verify that LinUCB has the same form of UCB.

3 Regret Analysis

We have so far directly applied the "optimism in face of uncertainty" principle to derive the LinUCB algorithm. The final step is to prove a regret bound for this algorithm.

Theorem 1. *If* $\lambda \geq 1$*, then the pseudo-regret of LinUCB is bounded as*

$$\bar{\mathcal{R}}_T \leq 2T\delta + \beta_T \sqrt{8dT \ln\left(1 + \frac{T}{\lambda d}\right)}.$$

Setting $\lambda = 1$ *and* $\delta = 1/T$ *leads to* $\bar{\mathcal{R}}_T = \mathcal{O}(d \ln(T/d) \sqrt{T})$.

Proof. Since $\langle a_t - a_t^\star, \theta^\star \rangle \leq |\langle a_t, \theta^\star \rangle| + |\langle a_t^\star, \theta^\star \rangle| \leq 2$, if suffices to show that under the event $\theta^\star \in \Theta_t$ for all t (which happens with probability $1 - \delta$ according to Lemma 1), we have $\bar{\mathcal{R}}_T \leq \beta_T \sqrt{8dT \ln\left(1 + \frac{T}{\lambda d}\right)}$. Indeed, notice that for any a,

$$\left| \left\langle a, \theta^\star - \widehat{\theta}_t \right\rangle \right| \leq \left\| \theta^\star - \widehat{\theta}_t \right\|_{M_t} \|a\|_{M_t^{-1}} \leq \beta_t \|a\|_{M_t^{-1}}.$$

Therefore with $a = a_{t+1}$ and $a = a_{t+1}^\star$, we have

$$\langle a_{t+1}, \theta^\star \rangle \leq \left\langle a_{t+1}, \widehat{\theta}_t \right\rangle + \beta_t \|a_{t+1}\|_{M_t^{-1}}$$

and

$$\left\langle a_{t+1}^\star, \theta^\star \right\rangle \geq \left\langle a_{t+1}^\star, \widehat{\theta}_t \right\rangle - \beta_t \left\| a_{t+1}^\star \right\|_{M_t^{-1}} \geq \left\langle a_{t+1}, \widehat{\theta}_t \right\rangle - \beta_t \|a_{t+1}\|_{M_t^{-1}}$$

where the last inequality is by the algorithm. Combining the above two inequalities we have

$$\left\langle a_{t+1} - a^\star_{t+1}, \theta^\star \right\rangle \leq 2\beta_t \left\| a_{t+1} \right\|_{M_t^{-1}} \leq 2\beta_T \left\| a_{t+1} \right\|_{M_t^{-1}}.$$

With the trivial bound shown at the beginning and $\beta_T \geq \sqrt{\lambda} \geq 1$, we have

$$\left\langle a_{t+1} - a^\star_{t+1}, \theta^\star \right\rangle \leq 2\beta_T \min\{1, \left\| a_{t+1} \right\|_{M_t^{-1}}\}.$$

and therefore by Cauchy-Schwarz inequality and $\min\{1, x\} \leq 2\ln(1+x)$, the regret is bounded by

$$\sqrt{T\sum_{t=1}^{T}(\langle a_t - a^\star_t, \theta^\star\rangle)^2} \leq \beta_T \sqrt{4T\sum_{t=1}^{T}\min\left\{1, \|a_t\|^2_{M_{t-1}^{-1}}\right\}}$$
$$\leq \beta_T \sqrt{8T\sum_{t=1}^{T}\ln\left(1+\|a_t\|^2_{M_{t-1}^{-1}}\right)} = \beta_T \sqrt{8T\ln\prod_{t=1}^{T}\left(1+\|a_t\|^2_{M_{t-1}^{-1}}\right)}.$$

Next by the fact that $\det(AB) = \det(A)\det(B)$ and $I + vv^\top$ has only two eigenvalues 1 and $1 + \|v\|_2^2$, we have

$$\det(M_T) = \det(M_{T-1} + a_T a_T^\top) = \det(M_{T-1}^{\frac{1}{2}}(I + M_{T-1}^{-\frac{1}{2}} a_T a_T^\top M_{T-1}^{-\frac{1}{2}}) M_{T-1}^{\frac{1}{2}})$$
$$= \det(M_{T-1})\det(I + M_{T-1}^{-\frac{1}{2}} a_T a_T^\top M_{T-1}^{-\frac{1}{2}}) = \det(M_{T-1})(1 + \|a_T\|^2_{M_{T-1}^{-1}})$$
$$= \cdots = \det(M_0)\prod_{t=1}^{T}\left(1 + \|a_t\|^2_{M_{t-1}^{-1}}\right).$$

It thus remains to show $\ln\frac{\det(M_T)}{\det(M_0)} \leq d\ln\left(1 + \frac{T}{\lambda d}\right)$. This is because by AM-GM inequality,

$$\det(M_T) \leq \left(\frac{\mathrm{TR}(M_T)}{d}\right)^d = \left(\frac{\lambda d + \sum_{t=1}^{T}\mathrm{TR}(a_t a_t^\top)}{d}\right)^d = \left(\lambda + \frac{\sum_{t=1}^{T}\mathrm{TR}(a_t^\top a_t)}{d}\right)^d \leq \left(\lambda + \frac{T}{d}\right)^d.$$

This finishes the proof together with $\det(M_0) = \lambda^d$. □

This regret bound for LinUCB has a linear dependence on the dimension d, which was shown to be optimal in the worst case (but suboptimal for the special case of multi-armed bandit where the dependence should be $\sqrt{d}$).

As a final remark, note that just as UCB, one can also derive a "gap-dependent" regret bound for LinUCB. Specifically, let the minimal suboptimal gap be

$$\Delta = \min_{t\in[T]}\min_{a\in A_t : \langle a - a^\star_t, \theta^\star\rangle > 0} \langle a - a^\star_t, \theta^\star\rangle.$$

Then one has either $\Delta \leq \langle a_t - a^\star_t, \theta^\star\rangle$ or $\langle a_t - a^\star_t, \theta^\star\rangle = 0$, and therefore

$$\bar{\mathcal{R}}_T \leq \frac{1}{\Delta}\sum_{t=1}^{T}(\langle a_t - a^\star_t, \theta^\star\rangle)^2,$$

where the last summation can be upper bounded in the exact same way as the proof above. This shows a regret of order $\mathcal{O}\left(\frac{(d\ln(T/d))^2}{\Delta}\right)$.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, 2011.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.