Lecture 3: Stochastic Bandits

Akshay Krishnamurthy
akshay@cs.umass.edu

February 7, 2022

1 Introduction

Recap on online learning. In the last lecture, we introduce the online learning protocol and saw some algorithms for making sequential predictions. We concluded the lecture by discussing the multi-armed bandit problem as a simple framework for developing algorithms that can explore. We saw the EXP3 algorithm, which balances exploration and exploitation to achieve $O(\sqrt{AT \log A})$ regret in the adversarial setting.

Note that the bandit protocol we studied does not require generalization anymore. Indeed in our formulation, there are no features whatsoever and, while features could be introduced, we are competing with the best fixed action rather than best "feature-dependent" policy. So we should think of algorithms like Exp3 as exhibiting just the exploration capability. (There are extensions that can handle generalization but we may not discuss them.)

To work our way towards algorithms that can generalize and explore, today we will first study the *stochastic multi-armed bandit* protocol, which is a simpler setting than we saw last time, so it still doesn't require generalization. Here we will develop some statistical algorithms that showcase the algorithmic principle of *optimism* and some new tools for incorporating the generalization capability.

2 Stochastic Multi-armed bandits

Optimism in the face of uncertainty provides an intuitive way to think about exploration. The idea is that, if we don't know everything about the environment, we should hope that it is most favorable to us and act accordingly. We will introduce this idea in the reward-formulation of multi-armed bandits. Working with rewards is not fundamentally different from the loss formulation, but somewhat more in line with the reinforcement learning literature.

The stochastic multi-armed bandit setting follows the same protocol as we saw in the last lecture, except that each arm a is associated with a distribution $\nu(a)$ and in each round the rewards are generated by sampling $r_t(a) \sim \nu(a)$. We write $\mu(a) = \mathbb{E}[r(a)]$ to be the mean and let us continue to assume that $r_t(a) \in [0, 1]$. (Note that this is a special case of the previous setup, after translating losses to rewards.)

Since the rewards are stochastic it is natural to try to estimate their means $\mu(a)$. The optimism principle is that: by a concentration argument we can be confident that each true mean $\mu(a)$ is sandwiched in the interval $\hat{\mu}(a) \pm \mathrm{conf}(a)$ for some carefully defined notion of conf. Acting optimistically means choosing the action that maximizes $\mathrm{argmax}_a \hat{\mu}(a) + \mathrm{conf}(a)$. Assuming our confidence intervals are correct, this is the best reward we could hope to achieve. See Figure 2 for an illustration.

The *upper confidence bound* algorithm, or UCB, implements exactly this strategy. Let $N_t(a) = \sum_{\tau=1}^{t} \mathbf{1}\{a_\tau = a\}$ denote the number of times we have pulled arm a up to and including round t. Then we can define

$$\hat{\mu}_t(a) = \frac{1}{N_t(a)} \sum_{\tau=1}^{t} r_t(a) \mathbf{1}\{a_t = a\}, \qquad \mathrm{conf}_t(a) = \sqrt{\frac{\log(2AT/\delta)}{N_t(a)}},$$

where $\delta \in (0, 1)$ is our failure probability parameter. The choice of conf arises from a concentration and union bounding argument. Then, after pulling each arm once, we set

$$a_t = \operatorname*{argmax}_a \hat{\mu}_{t-1}(a) + \mathrm{conf}_{t-1}(a).$$

Figure 1: An illustration of the UCB algorithm, showing how every time we play a suboptimal action we acquire information. In the round on the left, arm a is the UCB but it is suboptimal. It will be chosen and the corresponding confidence interval will shrink substantially. In this case the optimal arm $a^\star$ becomes the UCB afterwards.

To understand why this algorithm might work well it is helpful to consider two cases. First, if the arm that we play is optimal, meaning $a_t = a^\star$, then we incur no regret. The other case is when $a_t \neq a^\star$. If this happens, the confidence interval for a_t must be quite large, because the true mean $\mu(a_t)$ is in the confidence interval and this value is smaller that $\mu(a^\star)$. Thus, when we play a_t, even though we incur some regret, we will shrink the confidence interval for a_t substantially. Intuitively, every time we incur regret, we learn a lot about the unknown means.

The next theorem formalizes this intuition and provides a regret bound for the UCB algorithm.

Theorem 1. *UCB ensures that, with probability* $\geq 1-\delta$*:* $\max_a T\mu(a) - \sum_t \mu(a_t) \leq O\left(\sqrt{AT\log(AT/\delta)}\right)$.

Proof. First, by a union bound and a (slightly subtle) concentration argument, we have with probability $1-\delta$:

$$\forall t \in [T], a \in [A] : |\hat{\mu}_t(a) - \mu(a)| \leq \mathrm{conf}_t(a).$$

Assuming this holds, we use the optimistic regret decomposition. Let $a^\star = \mathrm{argmax}_a\, \mu(a)$. Then for $t > A$:

$$\begin{aligned}
\mu(a^\star) - \mu(a_t) &= \mu(a^\star) - \hat{\mu}_{t-1}(a^\star) + \hat{\mu}_{t-1}(a^\star) - \hat{\mu}_{t-1}(a_t) + \hat{\mu}_{t-1}(a_t) - \mu(a_t) \\
&\leq \mathrm{conf}_{t-1}(a^\star) + \hat{\mu}_{t-1}(a^\star) - \hat{\mu}_{t-1}(a_t) + \mathrm{conf}_{t-1}(a_t) \\
&\leq 2 \cdot \mathrm{conf}_{t-1}(a_t)
\end{aligned}$$

where the second step uses the confidence bound and the last step uses the selection rule. Thus, the regret is

$$\sum_t (\mu(a^\star) - \mu(a_t)) \leq A + 2\sum_{t=A+1}^{T} \mathrm{conf}_{t-1}(a_t) \leq A + 2\sqrt{\log(2AT/\delta)} \cdot \sum_{t=A+1}^{T} \sqrt{\frac{1}{N_{t-1}(a_t)}}$$

To bound this last sum, we need to use the fact that $N_t(a_t) \geq 1$ (since we pull each arm once to start) and $N_t(a_t) = N_{t-1}(a_t) + 1$ so that all the counts are increasing. Then

$$\sum_{t=A+1}^{T} \sqrt{\frac{1}{N_{t-1}(a_t)}} = \sum_a \sum_{j=1}^{N_T(a)} \sqrt{\frac{1}{j}} \leq 2\sum_a \sqrt{N_T(a)} \leq 2\sqrt{AT}$$

This second to last inequality follows from the bound $\sum_{i=1}^{n} 1/\sqrt{i} \leq 2\sqrt{n}$ which can be proven by induction, while the last inequality uses Cauchy-Schwarz. Essentially, the worst case allocation is uniform over all the arms. □

Note that optimism plays a critical role in the proof, by allowing us to (roughly speaking) replace the $\mathrm{conf}(a^\star)$ term with a second $\mathrm{conf}(a_t)$ term. This is crucial because $\mathrm{conf}(a^\star)$ will not necessarily shrink in each round, but $\mathrm{conf}(a_t)$ certainly will. The fact that the regret is bounded by our own confidence interval also formalizes the intuition that when we incur regret, we must learn a lot.

While we will not cover this in the course, we note that one can obtain stronger guarantees, known as instance-dependent or gap-dependent bounds for the UCB strategy.

3 Stochastic linear bandits

Since the UCB strategy seems quite flexible, it is a natural algorithm to try in many other exploration settings. Here we will consider a setting that does requires some amount of generalization, known as *linear bandits*.

The protocol is the same as in stochastic multi-armed bandits, except we will consider a large, potentially infinite, set of actions $\mathcal{A}$ where each action a is associated with a feature vector $\phi(a) \in \mathbb{R}^d$. Instead of parametrizing the reward distribution for each action separately, we assume there is a single vector $\theta^\star \in \mathbb{R}^d$ such that

$$\forall a \in \mathcal{A} : r(a) \sim \mathcal{N}(\langle \phi(a), \theta^\star \rangle, \sigma^2).$$

In particular, we have $\mathbb{E}[r(a)] = \langle \phi(a), \theta^\star \rangle$ which is known as *linear realizability*. Note that different noise distributions can be considered, but Gaussian noise is easy to work with while retaining all of the main ideas.

In some sense this is a special case of the stochastic MAB protocol, so we can simply run UCB while ignoring all of the linear structure. This will achieve $\sqrt{|\mathcal{A}|T\log(|\mathcal{A}|)}$ regret, which could be quite bad if the number of actions is large. The hope in linear bandits is to exploit the additional structure to replace all instances of $|\mathcal{A}|$ with the feature dimension d. Indeed we will see how to achieve $O(d\sqrt{T\log(d)})$ regret, which allows us to scale to large action spaces and captures some notion of generalization.

Since we are in a stochastic setting we can try to instantiate a UCB strategy. We need to answer two questions: (1) How do we do estimation? (2) How do we construct the confidence intervals? The answer to both of these questions will come from basic properties of linear regression.

Intuition from linear regression. Since we have linear structure, the most natural estimator is via linear regression. To build intuition, it is helpful to review the analysis of "fixed design" linear regression. Suppose we have a dataset $\{(\phi_i, r_i)\}_{i=1}^n$ of feature vectors and their associated rewards where $\phi_i \in \mathbb{R}^d$ and $r_i \sim \mathcal{N}(\langle \phi_i, \theta^\star \rangle, \sigma^2)$. Then we can form an estimate $\hat{\theta}$ by solving the following problem:

$$\hat{\theta} \leftarrow \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \left(\langle \phi_i, \theta \rangle - r_i\right)^2, \qquad \hat{\theta} = (\Phi^\top \Phi)^{-1} \Phi^\top R,$$

where we have written the estimator in a closed form using the matrix notation: $\Phi \in \mathbb{R}^{n \times d}$ has rows corresponding to the feature vectors and $R \in \mathbb{R}^n$ has the rewards. (For now let us assume that $n \geq d$ and $\Phi^\top \Phi$ is invertible.)

By realizability, we can write $R = \Phi\theta^\star + Z$ where $Z \in \mathbb{R}^n$ is a vector of $\mathcal{N}(0, \sigma^2)$ entries. Therefore

$$\hat{\theta} - \theta^\star = (\Phi^\top \Phi)^{-1} \Phi^\top \Phi \theta^\star + (\Phi^\top \Phi)^{-1} \Phi^\top Z - \theta^\star = (\Phi^\top \Phi)^{-1} \Phi^\top Z.$$

The cleanest way to control the last term is to express the error in a certain norm induced by the features. For a positive definite matrix M let $\|x\|_M^2 = x^\top M x$ be the *Mahalanobis norm*. Then taking expectation over the noise vector Z gives:

$$\mathbb{E}_{Z \sim \mathcal{N}(0,\sigma^2)} \left[\|\hat{\theta} - \theta^\star\|_{\Phi^\top \Phi}^2 \right] = \mathbb{E}_{Z \sim \mathcal{N}(0,\sigma^2)} \left[\|(\Phi^\top \Phi)^{-1} \Phi^\top Z\|_{\Phi^\top \Phi}^2 \right] = \mathbb{E}_{Z \sim \mathcal{N}(0,\sigma^2)} [Z^\top \Phi (\Phi^\top \Phi)^{-1} \Phi^\top Z] = \sigma^2 d \tag{1}$$

The last step here follows from the fact that we are projecting an n-dimensional gaussian vector onto a d dimensional subspace. Formally, by trace rotation $\mathbb{E}[Z^\top M Z] = \mathbb{E}\operatorname{tr}(MZZ^\top) = \sigma^2 \operatorname{tr}(M)$. Finally, by the Cauchy-Schwarz inequality for any feature ϕ

$$\mathbb{E}_{Z \sim \mathcal{N}(0,\sigma^2)} \left[\left\langle \phi, \hat{\theta} - \theta^\star \right\rangle \right] \leq \mathbb{E}_{Z \sim \mathcal{N}(0,\sigma^2)} \left[\|\phi\|_{(\Phi^\top \Phi)^{-1}} \|\hat{\theta} - \theta^\star\|_{\Phi^\top \Phi} \right] \leq \|\phi\|_{(\Phi^\top \Phi)^{-1}} \sqrt{\sigma^2 d}. \tag{2}$$

As we can see, the matrix $\Phi^\top \Phi = \sum_{i=1}^n \phi_i \phi_i^\top$, which is the covariance matrix of the features plays a critical role in the convergence of linear regression. The norm $\|\cdot\|_{\Phi^\top \Phi}$ is sometimes called the *data norm* and it measure how much "signal" we have in each direction. For example, if the features are the standard basis elements e_i, it is natural to expect that we have a better estimate of $\theta_i^\star$ for coordinates i that we have seen many times.

LinUCB. Although we cannot use exactly the above analysis, the intuition is very helpful when designing the algorithm. To avoid the requirement that the covariance matrix is invertible, we will use a regularized version of linear regression to form an estimate $\hat{\theta}_t$ at round t. Then we will use the data norm to form our confidence interval and the UCBs will be based on the inverse data norm, drawing inspiration from Eq. (2).

At round t we have data $\{(\phi(a_i), r_i)\}_{i=1}^{t-1}$ and we solve the *ridge regression problem*

$$\hat{\theta}_t \leftarrow \operatorname*{argmin}_{\theta} \sum_{i=1}^{t-1} (\langle \phi(a_i), \theta \rangle - r_i)^2 + \lambda \|\theta\|_2^2, \qquad \hat{\theta}_t = \Sigma_t^{-1} \sum_{\tau=1}^{t-1} \phi(a_\tau) r_\tau,$$

where $\Sigma_t = \lambda I + \sum_{i=1}^{t-1} \phi(a_i)\phi(a_i)^\top$ is the "regularized" feature covariance at time t. The confidence interval is:

$$\textsc{Ball}_t = \left\{ \theta : \|\theta - \hat{\theta}_t\|_{\Sigma_t} \le \beta \right\},$$

where we will set β roughly like $\sigma^2 d$. Then, the UCB strategy is

$$a_t = \operatorname*{argmax}_{a \in \mathcal{A}} \max_{\theta \in \textsc{Ball}_t} \langle \phi(a), \theta \rangle .$$

A similar strategy defines the confidence bound as $\mathrm{conf}_t(a) = \sqrt{\beta}\|\phi(a)\|_{\Sigma_t^{-1}}$ and maximizes $\left\langle \phi(a), \hat{\theta}_t \right\rangle + \mathrm{conf}_t(a)$.

For this algorithm we can prove the following theorem. Many variations of this result are possible.

Theorem 2. *Suppose that $\|\theta^\star\| \le W$ and $\forall a \in \mathcal{A} : |\langle \phi(a), \theta^\star \rangle| \le 1, \|\phi(a)\| \le B$. Set*

$$\lambda = \sigma^2/W^2, \quad \beta = \sigma^2 \left(2 + 4d \log\left(1 + \frac{TB^2W^2}{d} \right) + 8 \log(4/\delta) \right).$$

Then, there is a universal constant $C > 0$ such that, with probability at least $1 - \delta$ we have:

$$T \cdot \max_{a \in \mathcal{A}} \langle \phi(a), \theta^\star \rangle - \sum_t \langle \phi(a_t), \theta^\star \rangle \le C\sigma\sqrt{T} \left(d \log\left(1 + \frac{TB^2W^2}{d\sigma^2} \right) + \log(4/\delta) \right).$$

Proof. The first claim is that with probability at least $1 - \delta$ we have

$$\forall t \in [T] : \theta^\star \in \textsc{Ball}_t.$$

We will not prove this claim here as it is quite technical, but you may refer to Chapter 6.3 of the RL theory monograph for a detailed proof. However, the choice of β scaling as $d\sigma^2$ should be somewhat predictable based on Eq. (1) while a dependence on $\log(T)$ and $\log(1/\delta)$ arise from a union bound over time.

Assuming the above is true, we can proceed with the optimistic regret decomposition. At round t we have

$$\begin{aligned}
\langle \phi(a^\star), \theta^\star \rangle - \langle \phi(a_t), \theta^\star \rangle &\le \max_{\theta \in \textsc{Ball}_t} \langle \phi(a^\star), \theta \rangle - \langle \phi(a_t), \theta^\star \rangle \\
&\le \max_{\theta \in \textsc{Ball}_t} \langle \phi(a_t), \theta \rangle - \langle \phi(a_t), \theta^\star \rangle \\
&= \max_{\theta \in \textsc{Ball}_t} \left\langle \phi(a_t), \theta - \hat{\theta}_t \right\rangle - \left\langle \phi(a_t), \theta^\star - \hat{\theta}_t \right\rangle \\
&\le 2\sqrt{\beta} \cdot \|\phi(a_t)\|_{\Sigma_t^{-1}}
\end{aligned}$$

The first inequality uses that $\theta^\star \in \textsc{Ball}_t$, while the second uses that a_t is the UCB action so it maximizes the "index" $\max_{\theta \in \textsc{Ball}_t} \langle \phi(a), \theta \rangle$. In the last step we use Cauchy-Schwarz just as we did to obtain Eq. (2). Actually since we assumed that mean rewards are in $[-1, 1]$ the per-round regret is at most 2, which is another bound we can use for free.

Thus, we can bound the cumulative regret by

$$\sum_t \langle \phi(a^\star), \theta^\star \rangle - \langle \phi(a_t), \theta^\star \rangle \le 2\sqrt{\beta} \sum_t \min\left(1, \|\phi(a_t)\|_{\Sigma_t^{-1}} \right) \le 2\sqrt{\beta T} \sqrt{\sum_{t=1}^{T} \min\left(1, \|\phi(a_t)\|_{\Sigma_t^{-1}}^2 \right)} \tag{3}$$

This is essentially the same argument we did for the few-actions case, except we have a different notion of confidence. For intuition, you can think of $\|\phi(a_t)\|_{\Sigma_t^{-1}}$ as analogous $\sqrt{1/N_t(a)}$ which we saw in the previous proof.

To finish the proof we need a more intricate potential function argument.

Lemma 3 (Elliptical potential lemma)**.** *Let $x_1, \ldots, x_T$ be a sequence of vectors with $\|x_t\|_2 \leq B$ and define $\Sigma_1 = \lambda I$, $\Sigma_{t+1} = \Sigma_t + x_t x_t^\top$. Then*

$$\sum_{t=1}^{T} \min(1, x_t^\top \Sigma_t^{-1} x_t) \leq d \log\left(1 + \frac{TB^2}{d\lambda}\right)$$

Proof. First we claim that $\min(1, x_t^\top \Sigma_t^{-1} x_t) \leq 2 x_t^\top \Sigma_{t+1}^{-1} x_t$, where we have shifted the index on the covariance matrix by one. The key here is to use the Sherman-Morrison-Woodbury formula for rank-one updates to a matrix inverse:

$$\begin{aligned} x_t^\top \Sigma_{t+1}^{-1} x_t = x_t^\top (\Sigma_t + x_t x_t^\top)^{-1} x_t &= x_t^\top \left(\Sigma_t^{-1} - \frac{\Sigma_t^{-1} x_t x_t^\top \Sigma_t^{-1}}{1 + \|x_t\|_{\Sigma_t^{-1}}^2}\right) x_t \\ &= \|x_t\|_{\Sigma_t^{-1}}^2 - \frac{\|x_t\|_{\Sigma_t^{-1}}^4}{1 + \|x_t\|_{\Sigma_t^{-1}}^2} = \frac{\|x_t\|_{\Sigma_t^{-1}}^2}{1 + \|x_t\|_{\Sigma_t^{-1}}^2} \end{aligned}$$

Now let us consider two cases. If $x_t^\top \Sigma_t^{-1} x_t \leq 1$, then we can lower bound the RHS by $\|x_t\|_{\Sigma_t^{-1}}^2 / 2$ immediately. If $x_t^\top \Sigma_t^{-1} x_t \geq 1$ then observe that the RHS is directly at least $1/2$ since $\frac{z}{1+z}$ is increasing in x.

Next, the concavity of the log-determinant function means that $\log\det(\Sigma_t) \leq \log\det \Sigma_{t+1} + \mathrm{tr}(\Sigma_{t+1}^{-1}(\Sigma_t - \Sigma_{t+1}))$ by a first-order Taylor approximation. This gives us a telescoping sum

$$\sum_{t=1}^{T} \min(1, x_t^\top \Sigma_t^{-1} x_t) \leq \sum_{t=1}^{T} x_t^\top \Sigma_{t+1}^{-1} x_t = \sum_{t=1}^{T} \mathrm{tr}(\Sigma_{t+1}^{-1}(\Sigma_{t+1} - \Sigma_t)) \leq \sum_{t=1}^{T} \log\det \Sigma_t - \log\det \Sigma_{t-1} = \log\left(\frac{\det \Sigma_{T+1}}{\det \Sigma_1}\right)$$

This last term can be bounded using the norm bound on x_t. Observe that we must bound $\det(I + \frac{1}{\lambda}\sum_t x_t x_t^\top)$. On one hand, we have $\mathrm{tr}(\lambda^{-1}\sum_t x_t x_t^\top) = \lambda^{-1}\sum_t \|x_t\|_2^2 \leq B^2 T/\lambda$, so the sum of the eigenvalues is at most $B^2 T/\lambda$. Then by applying the AM-GM inequality to the eigenvalues $(\sigma_1, \ldots, \sigma_d)$ of the matrix $\lambda^{-1}\sum_t x_t x_t^\top$, we have

$$\log\det\left(1 + \frac{1}{\lambda}\sum_t x_t x_t^\top\right) = d \log\left(\prod_{i=1}^{d}(1+\sigma_i)\right)^{1/d} \leq d \log\left(\frac{1}{d}\sum_{i=1}^{d}(1+\sigma_i)\right) \leq d \log\left(1 + \frac{TB^2}{d\lambda}\right) \qquad \square$$

Using Lemma 3 in Eq. (3) proves the theorem. $\square$

There are many variations on the stochastic linear bandit problem for which algorithms like LinUCB can achieve $\mathrm{poly}(d)\sqrt{T}$ regret bounds. One example is the "linear contextual bandit" problem where the feature vectors change from round to round, perhaps as different users come to the system, but linear realizability continues to hold. In this case we will learn to compete with the best policy, which chooses the best action for each user and is therefore able to generalize across users.