Lecture 18: Stochastic Bandits

Akshay Krishnamurthy
akshay@cs.umass.edu

November 9, 2017

1 Recap

Last time we talked about the nonstochatsic bandit problem which was a partial information version of our online learning problem. Here we studied situations where at each iteration t, the learner chooses an action a_t and suffers loss $\ell_t(a_t)$ which is the only thing the learner observes. We showed that the importance weighting trick can be plugging into any full information algorithm with a local-norm type regret bound, which lead to a $O(\sqrt{KT\log(K)})$-regret for EXP3.

Today we'll study an easier version of this problem where the losses are not generated by an adversary. This is the stochastic multi-armed bandit problem.

2 Stochastic Bandits

Here we'll be in essentially the same setup as before, except that each arm a is associated with a distribution ν_a supported on $[0,1]$ with mean μ_a. At every round the rewards $r(a) \sim \nu_a$ iid for all arms but otherwise the game is exactly the same as before. For each of T rounds, we choose an arm a_t, suffer loss $r_t(a_t)$, which we also observe. Here the regret is

$$\text{Reg}(T) = \max_{a\in[K]} \sum_{t=1}^T r_t(a) - r_t(a_t)$$

We'll also study the pseudo-regret, which is

$$\bar{\text{Reg}}(T) = \max_{a\in[K]} \mathbb{E}\sum_{t=1}^T r_t(a) - r_t(a_t) = \max_{a\in[K]} T\mu_a - \sum_{t=1}^T \mathbb{E}\mu(a_t)$$

Here the expectation accounts for all randomness in the problem and in the agent. Note that the index a_t is a random variable, since it may depend on all of the previous observations. Then since, as we have seen, $\mathbb{E}\max \geq \max\mathbb{E}$ the pseudo-regret is in general smaller than the expected regret. It doesn't really have a natural interpretation, but it is a quantity that is heavily studied in the literature.

One thing to note here is that since we are not facing an adversary, we can now be a deterministic algorithm over the space of arms. We do not have to lift to the space of distributions. Of course since the stochastic case is a special case of the adversarial one, we know that EXP3 already has an $O(\sqrt{KT\log(K)}$ expected regret bound here. However today we will study some deterministic algorithms.

3 Epsilon-Greedy strategies

A simple strategy, that maybe many of you have seen, is based on uniform exploration. At each round, with probability ϵ, we choose an arm uniformly at random, and with probability $1-\epsilon$, we play the arm with the empirically highest mean. This is called *ϵ-greedy exploration.* Something that is simpler to analyze is the *explore-first* strategy which, plays uniformly at random for the first n rounds, and then plays the empirically best arm for the remaining rounds.

Theorem 1. *The Explore-First algorithm, with* $n = T^{2/3}(K\log(K/\delta))^{1/3}$ *has, with probability at least* $1-\delta$

$$\textit{Reg}(T) \le O((K\log(K))^{1/3}T^{2/3})$$

Proof Sketch. If we do n rounds of exploration and use the importance weighting reward estimator then using Hoeffding's inequality, we can prove

$$|\hat{\mu}(a) - \mu(a)| \le \sqrt{\frac{K\log(K/\delta)}{n}}$$

simultaneously for all $a \in [K]$, with probability $1-\delta$. Using the usual ERM analysis, this means that $\mu(\hat{a}) \ge \mu(a^\star) - 2\sqrt{K\log(K/\delta)/n}$. So our regret is

$$n + 2T\sqrt{\frac{K\log(K/\delta)}{n}}.$$

If we set $n = T^{2/3}(K\log(K/\delta))^{1/3}$ we get the desired bound. Technically one more deviation bound is required on $\sum_{t=n+1}^{T} r_t(a^\star) - r_t(\hat{a})$ but it is lower order. □

Essentially the same thing can be shown for ϵ-greedy strategies. However, since we know that EXP3 can get $\sqrt{KT\log(K)}$ regret, we know that this is suboptimal. Nevertheless, ϵ-greedy approaches can be quite effective in practice, and also in some cases are the best thing we currently know how to do. Getting to $\sqrt{T}$-type regret requires *adaptive exploration*, which roughly means localizing your exploration around the optimal arm, rather than doing something uniform. In some problems this can be hard, so ϵ-greedy is what we resort to.

4 Upper Confidence Bound Algorithms

The popular algorithm that people use for bandit problems is known as UCB for Upper-Confidence Bound. It uses a principle called "optimism in the face of uncertainty," which broadly means that if you don't know precisely what the environment is, make your decisions as if it were the best case for you. One way to think about this, which sometimes appears formally, is that when you make decisions optimistically, either you made a good decision, or you learned a lot. We'll see how this appears in the UCB analysis.

Let $N_t(a) = \sum_{i=1}^{t} \mathbf{1}\{a_i = a\}$ be the number of times you have pulled arm a up to and including round t. Define the empirical mean $\hat{\mu}_t(a) = \frac{1}{N_t(a)}\sum_{i=1}^{t} r_i(a)\mathbf{1}\{a_i = a\}$. Then the UCB algorithm is as follows. Start by pulling each arm once, and then at round t, the algorithm pulls

$$a_t = \operatorname*{argmax}_{a\in[K]} \hat{\mu}_{t-1}(a) + \sigma_{t-1}(a)$$

where $\sigma_{t-1}(a)$ is a confidence interval term that we'll set depending on $N_{t-1}(a)$. Intuitively it will be $O(\sqrt{1/N_{t-1}(a)})$, which is what you would get if you used Hoeffding's inequality on the empirical mean.

4.1 Distribution-Independent Analysis.

Theorem 2. *With* $\sigma_t(a) = \sqrt{\frac{\log(2KT/\delta)}{2N_t(a)}}$ *the regret of UCB is, with probability at least* $1-\delta$*,*

$$\textit{Reg}(T) \le O(\sqrt{KT\log(KT/\delta)})$$

Proof. The first step is to build the confidence intervals so that they trap the true mean with high probability. By Hoeffding's inequality and a union bound over all arms and all time T, with probability at least $1-\delta$

$$|\hat{\mu}_t(a) - \mu_t(a)| \le \sqrt{\frac{\log(2KT/\delta)}{2N_t(a)}}$$

Let us condition on this high probability event and proceed with the proof. The regret incurred in the first K rounds is at most K. Then, for any round $t > K$, our regret on that round is,

$$\begin{aligned}\mu(a^\star) - \mu(a_t) &= \mu(a^\star) + \hat{\mu}_{t-1}(a^\star) - \hat{\mu}_{t-1}(a^\star) - \mu(a_t) + \hat{\mu}_{t-1}(a_t) - \hat{\mu}_{t-1}(a_t)\\ &\le \sigma_{t-1}(a^\star) + \hat{\mu}_{t-1}(a^\star) + \hat{\mu}_{t-1}(a_t) + \sigma_{t-1}(a_t)\\ &\le 2\sigma_{t-1}(a_t)\end{aligned}$$

The first inequality uses the deviation bound both for $a^\star$ and for a_t. The second uses the decision rule of the algorithm, which implies that,

$$\hat{\mu}_{t-1}(a^\star) + \sigma_{t-1}(a^\star) \le \hat{\mu}_{t-1}(a_t) + \sigma_{t-1}(a_t)$$

Thus the total regret is,

$$\mathrm{Reg}(T) \le K + 2\sum_{t=K+1}^{T} \sigma_{t-1}(a_t) = K + \sqrt{2\log(KT/\delta)} \sum_{t=K+1}^{T} \sqrt{\frac{1}{N_{t-1}(a_t)}}$$

Now we have to bound this last term. We need to use that fact that if $a_t = a$ then $N_t(a_t) = N_{t-1}(a_t) + 1$, so that all of the terms are non-increasing. We also have to use that $N_K(a) = 1$ for all $i \in [K]$, which holds because we pull each arm once in the first K rounds. Thus we get,

$$\sum_{t=K+1}^{T} \sqrt{\frac{1}{N_{t-1}(a_t)}} = \sum_{a=1}^{K} \sum_{j=1}^{N_T(a)} \sqrt{\frac{1}{j}} \le \sum_{a=1}^{K} 2\sqrt{N_T(a)}.$$

The last inequality ($\sum_{i=1}^{n} \sqrt{1/i} \le 2\sqrt{n}$) can be proved by induction. Now finally, we know that $\sum_{i=1}^{K} N_T(a) \le T$. This is a concave constrained maximization, and it can be solved analytically. It turns out the worst case allocation is when $N_T(a) = T/K$ for each i, which gives the regret bound,

$$\mathrm{Reg}(T) \le K + \sqrt{8KT\log(KT/\delta)}.$$

□

Corollary 3. *UCB with* $\sigma_t(a) = \sqrt{\frac{\log(2KT^2)}{2N_t(a)}}$ *has pseudo-regret bound*

$$\bar{Reg}(T) = O(\sqrt{KT\log(KT)})$$

Proof. The bounds from above holds whenever the deviation bound holds. Trivially the regret is bounded by T otherwise so in total we get,

$$\bar{R}_n \le K + \sqrt{8Kn\log(Kn/\delta)} + \delta T$$

If we set $\delta = 1/T$ this bound becomes $O(\sqrt{KT\log(KT)})$. □

Modulo logarithmic factors, this is essentially the optimal regret achievable in the worst case. I like this analysis because you can see the exploration-exploitation tradeoff at work here. The regret is directly related to the deviation bounds, so a sharper deviation bounds give you better regret.

To see the optimism at work, notice that the instantaneous regret can only be big if $\sigma_{I_t,t-1}$ is big. Thus if you play a bad action, it must be because $\sigma_{I_t,t-1}$ was big, and by virtue of playing the action, you decrease $\sigma_{I_t,t-1}$ substantially. This captures the sentiment, either you play a good action, or you learn a lot.

4.2 Distribution-Dependent Analysis.

Now we'll prove an instance specific (or distribution dependent) bound. We need a new definition. Let,

$$\Delta(a) = \mu^\star - \mu(a)$$

be the suboptimality of arm a.

Theorem 4. *For $\sigma_t(a) = \sqrt{\frac{\log(t+1)}{N_t(a)}}$ the regret of the UCB algorithm is at most,*

$$\bar{Reg}(T) \leq \sum_{a:\Delta(a)>0} \left(\frac{4\log T}{\Delta(a)} + 8\Delta(a) \right)$$

This is called a distribution dependent bound, since the RHS depends on the gaps $\Delta(a)$ which are problem specific. Note that this bound has a much better dependence on T, of $O(\log(T))$ instead of $O(\sqrt{T})$, but if the gaps are extremely small, then the bound will not be that good.

Proof. First, we show that if $a_t = a$ then one of the following three inequalities must be true,

$$\begin{aligned} \hat{\mu}_{t-1}(a^\star) + \sigma_{t-1}(a^\star) &\leq \mu(a^\star) \\ \hat{\mu}_{t-1}(a) - \sigma_{t-1}(a) &> \mu(a) \\ \Delta(a) &\leq 2\sigma_{t-1}(a) \end{aligned}$$

Intuitively, if $\hat{\mu}(a)$ is small, $\hat{\mu}(a^\star)$ is big, and the gap is big then there is no way that $a_t = a$. Or formally, suppose that all three are false. Then

$$\hat{\mu}_{t-1}(a^\star) + \sigma_{t-1}(a^\star) > \mu^\star = \mu(a) + \Delta(a) \geq \mu(a) + 2\sigma_{t-1}(a) \geq \hat{\mu}_{t-1}(a) + \sigma_{t-1}(a)$$

which contradicts the fact that $a_t = a$.

Now we can bound the number of times we pull arm a. First, let us crudely solve for $N_{t-1}(a)$ in the inequality related Δ_a and $\sigma_{t-1}(a)$

$$\Delta_a \leq 2\sqrt{\frac{\log(T)}{N_{t-1}(a)}} \Rightarrow N_{t-1}(a) \geq \frac{4\log(T)}{\Delta_a^2}$$

So setting $u = \lceil 4\log(T)/\Delta(a)^2 \rceil$ we can bound

$$\begin{aligned} \mathbb{E}N_T(a) = \sum_{t=1}^{T} \mathbb{E}\mathbf{1}\{a_t = a\} &\leq \sum_{t=1}^{T} \mathbb{E}\mathbf{1}\{a_t = a \wedge N_{t-1}(a) \leq u\} + \mathbb{E}\mathbf{1}\{a_t = a \wedge N_{t-1}(a) > u\} \\ &\leq u + \sum_{t=1}^{T} \mathbb{P}[\hat{\mu}_{t-1}(a) - \sigma_{t-1}(a) > \mu(a)] + \mathbb{P}[\hat{\mu}_{t-1}(a^\star) + \sigma_{t-1}(a^\star) \leq \mu(a^\star)] \end{aligned}$$

Now we just need to upper bound the probability terms. They are identical so let's just look at the first one

$$\mathbb{P}[\hat{\mu}_{t-1}(a) - \sigma_{t-1}(a) > \mu(a)] \leq \exp\left(-2N_{t-1}(a)\sigma_{t-1}^2(a)\right) \leq \exp(-2\log(t)) \leq \frac{1}{t^2}$$

So we get

$$\mathbb{E}N_T(a) \leq u + 2\sum_{t=1}^{T} \frac{1}{t^2} \leq u + 4$$

The last thing is to use a useful decomposition of the pseudoregret

$$\bar{\text{Reg}}(T) = T\mu(a^\star) - \sum_{t=1}^{t} \mathbb{E}\mu(a_t) = T\mu(a^\star) - \sum_{a} \mathbb{E}N_T(a)\mu(a) = \sum_{a} \mathbb{E}N_T(a)\Delta(a) \leq \sum_{a} \frac{4\log(T)}{\Delta_a} + 4\Delta_a.$$

This proves the theorem. □