Lecture 17: Nonstochastic Bandits

Akshay Krishnamurthy
akshay@cs.umass.edu

November 9, 2017

1 Recap

For the last couple of lectures we have been studying online learning, where we repeatedly make predictions and suffer some loss, and the goal is to achieve low regret against a set of comparators. Specifically we have studied the *experts setting*, which proceeds for T rounds and in each round t we: (1) Play distribution $p_t \in \Delta([K])$, (2) Recieve loss $\ell_t \in [0,1]^K$, (3) Suffer loss $\langle p_t, \ell_t \rangle$. The regret here is measured as

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \langle p_t, \ell_t \rangle - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)$$

Today we will study a partial feedback version, where the full loss ℓ_t is not revealed to the learner.

2 Adversarial Bandits

In the *bandit setting*, the main difference is that rather than observe the entire loss vector ℓ_t, we just observe that coordinate of the loss corresponding to the arm that we choose. Or in more detail: (1) Choose distribution $p_t \in \Delta([K])$, (2) Sample $a_t \sim p_t$, suffer loss $\ell_t(a_t)$, (3) Observe $\ell_t(a_t)$. Now we measure the regret as

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \ell_t(a_t) - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)$$

For today let us just consider an *oblivious adversary* that sets all the losses before hand, but of course can simulate everything the algorithm will do. In this case, the comparator term is non-random, but our loss is a random variable. We also often care about the expected regret

$$\mathbb{E}\mathrm{Reg}(T) = \sum_{t=1}^{T} \mathbb{E}_{a_t \sim p_t}[\ell_t(a_t)] - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a) = \mathbb{E} \sum_{t=1}^{T} \langle p_t, \ell_t \rangle - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)$$

Of course p_t is also a random variable since it depends on the previous a_is, but bounding this is clearly adequate.

Importance weighting. The main trick to analyzing the bandit setting is to use *importance weights* to construct an unbiased estimate of the loss ℓ_t. Specifically, if at round t we play $a_t \sim p_t$ then we observe $\ell_t(a_t)$ and we construct the loss estimate

$$\tilde{\ell}_t(a) = \ell_t(a_t) \frac{\mathbf{1}\{a_t = a\}}{p_t(a)}$$

Such estimates have a number key properties which are extremely useful. Importance weights have a number of applications outside of bandit optimization, for example they are used in reinforcement learning, MCMC and probabilistic inference, and in several causal inference problems. You might also see them called Inverse Propensity Scores or IPS.

Lemma 1. *The importance weight estimate $\tilde{\ell}_t$ satisfies*

$$\mathbb{E}_{a_t \sim p_t} \tilde{\ell}_t(a) = \ell_t(a) \qquad \mathbb{E}_{a_t \sim p_t} \tilde{\ell}_t(a)^2 = \ell_t(a)^2 / p_t(a) \qquad \mathbb{E}_{a \sim p_t} \frac{1}{p_t(a)} = K$$

Proof. The claims are all require fairly straightforward calculations

$$\mathbb{E}_{a_t \sim p_t} \tilde{\ell}_t(a) = \sum_{a_t=1}^{K} p_t(a_t) \frac{\ell_t(a) \mathbf{1}\{a_t = a\}}{p_t(a)} = \ell_t(a)$$

$$\mathbb{E}_{a_t \sim p_t} \tilde{\ell}_t(a)^2 = \sum_{a_t=1}^{K} p_t(a_t) \frac{\ell_t(a)^2 \mathbf{1}\{a_t = a\}}{p_t(a)^2} = \frac{\ell_t(a)^2}{p_t(a)}$$

$$\mathbb{E}_{a \sim p_t} \frac{1}{p_t(a)} = \sum_{a=1}^{K} \frac{p_t(a)}{p_t(a)} = K$$

□

EXP3. Now that we understand the importance weight mechanism, the idea is to just plug these into the Hedge update to get a bandit algorithm. Specifically, starting with $w_1 = \mathbf{1} \in \mathbb{R}^K$, at each round t

1. Set $p_t \propto w_t$, sample $a_t \sim p_t$
2. Suffer loss $\ell_t(a_t)$, observe $\ell_t(a_t)$, construct $\tilde{\ell}_t$
3. Update $w_{t+1}(a) = w_t(a) \exp(-\eta \tilde{\ell}_t(a))$

We typically set $\eta = \sqrt{2 \log(K)/(TK)}$

Theorem 2. *EXP3 has expected regret bound*

$$\mathbb{E} Reg(T) \leq \sqrt{2KT \log(K)}$$

Proof. Since Hedge works against any adversary, we can just think of the adversary as playing the loss functions $\tilde{\ell}_t$, and applying the Hedge regret bound, we immediately get

$$\sum_{t=1}^{T} \langle p_t, \tilde{\ell}_t \rangle - \min_a \sum_{t=1}^{T} \tilde{\ell}_t(a) \leq \frac{\eta}{2} \sum_t \langle p_t, \tilde{\ell}_t^2 \rangle + \frac{\log(d)}{\eta}$$

Let us take expectation over this equation, specifically expectation over $a_t \sim p_t$. By the lemma, the first term on the LHS is the expected loss of the learner. For the second term

$$\mathbb{E} \min_a \sum_{t=1}^{T} \tilde{\ell}_t(a) \leq \min_a \mathbb{E} \sum_{t=1}^{T} \tilde{\ell}_t(a) = \min_a \sum_{t=1}^{T} \ell_t(a)$$

So if we take expectation over a_t then the LHS upper bounds the expected regret. Now we work on the RHS.

$$\mathbb{E}_{a_t \sim p_t} \langle p_t, \tilde{\ell}_t^2 \rangle = \sum_{a_t, a} p_t(a_t) p_t(a) \frac{\ell_t(a)^2 \mathbf{1}\{a_t = a\}}{p_t(a)^2} = \sum_a \ell_t(a)^2 \leq K$$

Hence we get

$$\mathbb{E} \mathrm{Reg}(T) \leq \frac{\eta}{2} TK + \frac{\log(K)}{\eta}.$$

The RHS here is optimized by setting $\eta = \sqrt{2 \log(K)/(TK)}$ and proves the result. □

This demonstrates the power of importance weighting, which can be used to convert a full information algorithm into a bandit algorithm. However note that we used a particular structure of the regret bound for the full information algorithm, namely the $\langle p_t, \ell_t^2 \rangle$ term. This term is called a *local norm*, since it can also be written as

$$\ell_t^T \mathrm{diag}(p_t) \ell_t = \|\ell_t\|^2_{\mathrm{diag}(p_t)}$$

which is more familiar to a Mahalanobis norm. In a sense, any full information algorithm with a local norm bound can be easily converted into a bandit algorithm using importance weighting.

Optimality. Now that we have seen how to derive bandit algorithms from full information ones, a natural question is whether the bound is optimal.

Theorem 3. *For any algorithm, there exists an adversary for which* $\mathbb{E}Reg(T) \geq \Omega(\sqrt{TK})$.

Proof Sketch. The idea is to consider a stochastic problem where the rewards for one arm (unknown to the learner) are drawn from $\mathrm{Ber}(1/2+\epsilon)$ while the rewards for all the other arms are drawn from $\mathrm{Ber}(1/2)$. Then, we compare the behavior of the algorithm on one of these instances, where the good arm is chosen randomly, against the behavior on an instance where all the arms are $\mathrm{Ber}(1/2)$. Then since one arm will be pulled less than T/K times, if ϵ is small enough the learner will not be able to distinguish between the situation where it is the best arm and all the arms are the same. This requires $\epsilon \leq \sqrt{K/T}$ and hence the learner will suffer $\sqrt{TK}$ regret. □

We will spend some time at the end of the course on how to prove lower bounds more formally. The point is that an algorithm for full information with local norms leads to an essentially optimal bandit algorithms.

Log barrier. In Homework 5 you will study another algorithm for the Experts setting based on the logarithmic barrier regularizer $R(p) = -\sum_{i=1}^{K} \log(p_i)$. In particular you will show a regret bound of

$$\mathrm{Reg}(T,u) \leq \eta \sum_{t=1}^{T} \langle p_t^2, \ell_t^2 \rangle + \frac{R(u)}{\eta}$$

The first term here is in some sense even better than it was for Hedge, but we have to worry about the second term, since it can be infinity if u is on the corner of the simplex. For this second term, we only apply the regret bound to comparators u of the form $u = (1-\epsilon)v + \epsilon/K$ in which case we get

$$R(u) = -\sum_{i=1}^{K} \log((1-\epsilon)v_i + \epsilon/K) \leq K\log(K) + K\log(1/\epsilon).$$

Now for any vector $v \in \Delta([K])$

$$\begin{aligned}
\mathrm{Reg}(T,v) &= \sum_{t=1}^{T} \langle p_t - v, \ell_t \rangle = \sum_t (1-\epsilon)\langle p_t - v, \ell_t \rangle + \epsilon \langle p_t - v, \ell_t \rangle \\
&= \sum_t (1-\epsilon)\langle p_t - v, \ell_t \rangle + \epsilon \langle p_t - \mathbf{1}/K, \ell_t \rangle + \epsilon \langle \mathbf{1}/K - v, \ell_t \rangle \\
&\leq \sum_t \langle p_t - u, \ell_t \rangle + \epsilon T = \mathrm{Reg}(T,u) + \epsilon T
\end{aligned}$$

In particular if we set $\epsilon = 1/T$ we pick up an additional regret of 1, which will be lower order. So now we have shown

$$\mathrm{Reg}(T) \leq \eta \sum_{t=1}^{T} \langle p_t^2, \ell_t^2 \rangle + K\log(TK)/\eta + 1$$

In the bandit setting, we can obtain the same bound on the expected regret but with $\tilde{\ell}_t$. The important point is

$$\mathbb{E}_{a_t \sim p_t} \langle p_t^2, \tilde{\ell}_t^2 \rangle = \sum_{a_t, a} p_t(a_t) p_t(a)^2 \frac{\ell_t(a)^2 \mathbf{1}\{a_t = a\}}{p_t(a)^2} \leq \langle p_t, \ell_t \rangle.$$

So instead of picking up TK here we are picking up the total loss of the learner. This can be used to derive a so-called *first-order* regret bound:

$$\sum_{t=1}^{T}\langle p_t, \ell_t\rangle - \sum_t \ell_t(a^\star) \le \eta \sum_t \langle p_t, \ell_t\rangle + K\log(TK)/\eta + 1$$

$$\Rightarrow \sum_{t=1}^{T}\langle p_t, \ell_t\rangle - \sum_t \ell_t(a^\star) \le \frac{1}{1-\eta}\left(\eta \underbrace{\sum_t \ell_t(a^\star)}_{\triangleq L^\star} + K\log(TK)/\eta + 1\right)$$

Now if we set $\eta = \min\{\sqrt{K\log(TK)/L^\star}, 1/2\}$ we get a bound of $O(\sqrt{KL^\star\log(TK)})$. This is like the $R(h^\star)$ type bounds we have seen in the homework several times, but in the bandit setting. The point is that obtaining such bounds in partial information settings is quite challenging and can require new algorithms, contrasting from the full information case where this can be fairly easy.

High probability bounds. So far we have analyzed the expected regret, but you might be wondering if we can say anything about the actual regret. This is the reason for high probability bounds. As above, obtaining a high probability bound unfortunately requires changing the algorithm. The reason is that while the second moment of the importance weights is well behaved (e.g., the $\langle p_t, \tilde{\ell}_t^2\rangle$ term), the range of these random variables could be quite poorly behaved, so we will not obtain good concentration.

The easiest way to change things is to mix in a little bit of uniform exploration. At round t, we instead choose $a_t \sim (1-\gamma)p_t + \gamma/d$ where γ is another parameter. This uniform exploration keeps the range in check and does not incur too much additional regret. Another small change is also required – rather than update using $\tilde{\ell}_t$, we update with an upper confidence bound, which looks like $\tilde{\ell}_t(a) + \frac{\alpha}{p_t(a)\sqrt{KT}}$. The algorithm enjoys the following guarantee

Theorem 4. *The EXP3.P algorihtm has a regret of $O(\sqrt{KT\log(KT/\delta)})$ with probability at least $1-\delta$.*

Another trick to obtain high probability bounds was developed recently by Gergely Neu. Rather than use unbiased importance weights, he decided to bias the estimates

$$\bar{\ell}_t(a) = \frac{\ell_t(a_t)\mathbf{1}\{a_t = a\}}{p_t(a) + \gamma}.$$

While this introduces bias, the fact that $\gamma > 0$ keeps the range of these random variables in check, and this suffices to obtain a high probability bound for vanilla EXP3, with these *implicit exploration* loss estimates.

3 Other bandit problems

As we saw other online learning problems besides the experts setting, there are also other bandit optimization problems. In bandit linear optimization we only observe $\langle w_t, \ell_t\rangle$, while in bandit convex optimization we only observe $f_t(w_t)$. Both settings require radically new techniques, since it is no longer clear how to design an unbiased loss estimator like we did in the standard bandit setting. In particular, regularization does not give us adequate exploration, and more intentional exploration is required.

In full information we saw that the linear case was the most challenge of the convex optimization cases, since we could always linearize the loss. While you can still linearize the loss in the bandit setting, you don't have access to the gradient, so you do not know how to do it. Indeed bandit convex optimization has only recently been adequately solved, and it is much more challenging than the linear case.