CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 26**

Online bandit problems

Lecturer: Alexander Rakhlin *Scribe: Fabian Wauthier*

In the past lectures we have addressed online learning in a setting where we observe the full function l_t. In this lecture we will look at online learning with limited feedback. We are interested in the question of how much feedback is needed in order to achieve low regret. Throughout this lecture, we will assume that the functions l_t are linear.

We have previously looked at the following game: For time $t = 1, 2, \dots, T$, we predict $x_t \in K \subset \mathbb{R}^n$. At each time step we observe $l_t \in \mathbb{R}^n$ and suffer $l_t \cdot x_t$. Then we take a gradient step. In this game we observe the whole loss function l_t after having chosen x_t and the last few lectures showed that a $\sqrt{T}$ regret can be achieved. We will now step away from the full information game and work in a limited feedback setting where functions l_t are linear, but we only observe the value $(l_t \cdot x_t)$ at each time step, rather than the whole function l_t. This lecture will address the question whether we can still achieve $\sqrt{T}$ regret if we are given only a single number. Before we begin, let us develop more motivation through some examples.

1. Online shortest path

   As discussed in the last lecture, the selection of a path in a directed acyclic graph $G = (E, V)$ can be formalised as a binary vector of length $|E|$, where a non-zero component indicates that the corresponding edge is used in this path. Here the graph G loosely corresponds to the set K. Let $\mathcal{P} \subseteq \{0,1\}^{|E|}$ be the subset of the binary hypercube corresponding to legal path vectors from a source to a sink. If arc i is associated with a delay l_t^i, then the total delay for a particular chosen path $x_t \in \mathcal{P}$ can be written as $l_t \cdot x_t$. In the limited feedback setting we will, after having chosen a path x_t, only observe the total delay $(l_t \cdot x_t)$. We are interested in choosing paths x_t in such a way that adversarially chosen delays on edges lead to low overall regret.

2. Stochastic multi-armed bandit

   In this problem, the set K is the simplex and the l_t are chosen stochastically. The component l_t^i is chosen i.i.d. from a fixed distribution with mean μ_i. The means are fixed throughout the game, and the l_t are not chosen adversarially. We wish to receive the largest gain throughout the game. The regret is given as

   $$R_T = \sum_{t=1}^{T} l_t \cdot x_t - \min_i \sum_{t=1}^{T} \mu_i \tag{1}$$

   Under the previous assumptions the formulation reduces to the multi-armed bandit problem, where one receives l_t^i reward for choosing arm i. The expected regret is the same for choosing fractional $x_t \subseteq K$ and for probabilistically choosing vertices of the simplex with some fractional probability.

3. Non-stochastic multi-armed bandit

   The setup for this problem is similar to the previous one, except that now l_t are chosen adversarially. We make no assumptions about an underlying distribution for l_t.

In the limited feedback formulation only one number $l_t \cdot x_t$ is given in response to a choice x_t. A first observation is that if the strategy is deterministic (i.e. x_t is one of the vertices of K), then the adversary

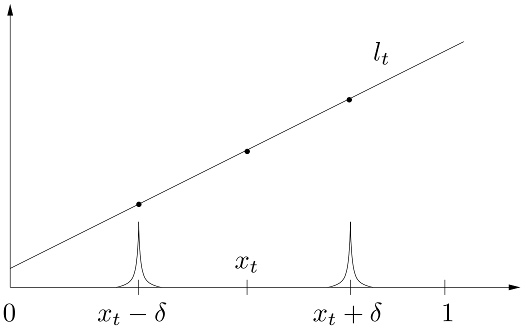

Figure 1: The choice y_t is computed from x_t by adding or subtracting a value of δ uniformly at random. The distribution of y_t is comprised of two delta functions and has mean x_t.

could mislead us by playing a function which has low value at our deterministic choice but is large elsewhere. If our choice is stochastic, then the adversary cannot always hide information about l_t from us.

So instead of predicting x_t directly, we will predict y_t, which is drawn from a distribution with mean x_t. To learn about the (linear) function l_t we want to estimate its slope by suitable queries y_t. Consider the following example in two dimensions. Suppose that $K = [0,1] \times \{1\}$ is a flat set, and that the adversary gives a vector-valued response $l_t = (\cdot,\cdot)^\top$. Here the first component of l_t corresponds to the slope of the linear function, and the second to an offset term (hence the fixed second dimension of K). By construction, our choice $x_t \in K$ is determined by its first component. Suppose the procedure tells us to predict x_t. Then define a randomised y_t as

$$y_t = \begin{cases} x_t + \delta & \text{w.p. } 1/2 \\ x_t - \delta & \text{w.p. } 1/2. \end{cases} \tag{2}$$

An estimator of the slope of the function l_t (its first component) is given by

$$\tilde{l}_t = \frac{l_t y_t}{\delta} \operatorname{sign}(y_t - x_t). \tag{3}$$

This estimator is unbiased, since

$$E\left(\tilde{l}_t\right) = \frac{1}{2}\left[\frac{l_t(x_t+\delta)}{\delta} - \frac{l_t(x_t-\delta)}{\delta}\right] = l_t \tag{4}$$

Although we start with one point x_t, we turn it into a randomised strategy for two points that to some extent recovers the function l_t of which we only see evaluations. This process is illustrated in Figure 1. Such a randomisation strategy also works in higher dimensions. First choose a coordinate i_t uniformly at random from $\{1,\ldots,n\}$. Pick the direction $\varepsilon_t = \pm 1$ with probability $1/2$ and let $y_t = x_t + \varepsilon_t \cdot \delta \cdot e_{i_t}$, where e_{i_t} is the i_t-th standard basis. We now let $\tilde{l}_t$ be defined as

$$\tilde{l}_t = n \frac{l_t y_t}{\delta} \varepsilon_t \cdot e_{i_t} \tag{5}$$

Here also, $E(\tilde{l}_t) = l_t$. Returning to the one-dimensional case, while the estimator in one dimension is unbiased, its variance depends on the denominator δ. We require for a “good” estimator of the slope that δ is as large as possible. Note in particular that if x_t lies near the boundaries at 0 or 1, then δ must be small and our estimator will be of high variance.

Once we have computed an estimate $\tilde{l}_t$, we will choose the mean of the next distribution x_{t+1} as

$$x_{t+1} = \text{argmin}_{x \in K} \eta \sum_{s=1}^{t} \tilde{l}_s x + R(x), \tag{6}$$

where $R(x)$ is a regulariser. This approach is called Follow The Regularized Leader. We recall that the key terms to control for low regret are $D_R(x_t, x_{t+1})$. Suppose that the regulariser R is chosen as $1/2||\cdot||^2$. Then we have $D_R(x_t, x_{t+1}) = D_{R^*}(0, \eta\tilde{l_t}) = \eta^2||\tilde{l}_t||^2$. Here, R is self-dual, so R^* is the same as R. Heuristically, whenever predicting a point which is close to the boundary, we can step away from the boundary by $O(1/\sqrt{T})$ without hurting the regret by more than $O(\sqrt{T})$. We can then guarantee that $\delta \geq O(1/\sqrt{T})$ so that the estimate of the slope $\tilde{l}_t$ scales as $1/\delta \approx \sqrt{T}$. Hence $D_R(x_t, x_{t+1}) \approx \eta^2 T$. This suggests that for the regret

$$\begin{aligned} R_T &\leq \eta^{-1} D_R(u, x_1) + \eta^{-1} T(\eta^2 T) && (7) \\ &\approx \eta^{-1} + \eta T^2 && (8) \\ &= O(T) && (9) \end{aligned}$$

The last step follows by choosing η to balance the two terms. We were hoping to get a regret on the order of $\sqrt{T}$, but under the previous assumption we only achieved a regret which is linear in T. The previous result can be improved by clever sampling choices for y_t only up to $O(T^{2/3})$ regret. For a long time it was unclear whether a regret on the order of $\sqrt{T}$ could be achieved. The key to proceeding was to exploit the additional freedom of choosing the regulariser R. Under a suitable choice of R it turns out that $O(\sqrt{T})$ regret is possible.

We will give a heuristic argument for the choice of R. We know that the problem occurs when x_t is close to the boundary, since then the variance of $\tilde{l}_t$ blows up as $1/d$, where d is the distance to the boundary. We also know that in order to get the desired regret, we need to control the divergence $D_{R^*}(0, \eta\tilde{l_t})$. This term is small if R^* is close to linear when $\tilde{l}_t$ is large. It turns out that the curvature in the primal corresponds to linearity in the dual. So heuristically, the primal R has to be relatively "curved" when $\tilde{l}_t$ is large, that is, when x_t is close to the boundary. If $\overline{H}$ is the Hessian of R^*, then

$$\begin{aligned} D_{R^*}(0, \eta\tilde{l}_t) &\approx \eta^2 \tilde{l}_t^\top \overline{H} \tilde{l}_t && (10) \\ &\overset{\text{want}}{=} O(\eta^2). && (11) \end{aligned}$$

The last equality must hold if we want to get $O(\sqrt{T})$ regret. So we need the linearity of R^* to kill the large magnitude l_t. Earlier we claimed that $\tilde{l}_t$ will scale approximately like $1/d$. It follows that in one dimension $\overline{H}$ should behave like d^2. The Hessian of R^* is the inverse Hessian of R, so then we see that the Hessian of R should behave like $1/d^2$. A function of d which has $1/d^2$ as its Hessian is $-\log(d)$. Hence the key property of R is that it behaves like the $-\log$ of the distance to the boundary. We have heuristically argued for key qualities of R. To understand how this function can be constructed for general sets, we appeal to the theory of interior point methods.

Definition. Self-concordant function

A function $R : K \to \mathbb{R}$ is self-concordant if it is a C^3 convex function such that

$$\left|D^3 R(x)[h,h,h]\right| \leq 2\left(D^2 R(x)[h,h]\right)^{3/2}, \tag{12}$$

where the third order differential is defined as

$$D^3 R(x)[h_1, h_2, h_3] = \frac{\partial^3}{\partial t_1 \partial t_2 \partial t_3}\bigg|_{t_1=t_2=t_3=0} R(x + t_1 h_1 + t_2 h_2 + t_3 h + 3) \tag{13}$$

Definition. θ-self concordant barrier

A θ-self concordant barrier is a self-concordant function with

$$|DR(x)[h]| \leq \theta^{1/2} \left[D^2R(x)[h,h]\right]^{1/2} \tag{14}$$

We now present some examples of θ-self concordant functions

1. If we have an n-dimensional ball, $B_n = \left\{x \in \mathbb{R}^n, \sum_i x_i^2 \leq 1\right\}$, then $R(x) = -\log(1 - ||x||^2)$ is 1-self concordant.
2. If we have a constraint $a^\top x \leq b$, then $R(x) = -\log(b - a^\top x)$. The constraint is a hyperplane and $R(x)$ is the negative log distance to the constraint.
3. Additionally, if R_1 is θ_1-self concordant and R_2 is θ_2-self concordant, then $R_1 + R_2$ is $(\theta_1 + \theta_2)$-self concordant.

We are now ready to show that for a self-concordant barrier R there is a suitable sampling strategy so that for the divergence $D_{R^*}(0, \eta\tilde{l}_t) = O(\eta^2)$. The latter will then imply that we can achieve $O(\sqrt{T})$ regret.

Define the local Euclidean geometry at $x \in K$ by $\langle g, h\rangle_x = g^\top \nabla^2 R(x) h$. This induces the norm at x to be $||h||_x = \sqrt{\langle h, h\rangle_x}$. That is, locally the space is stretched according to the Hessian at x. Define an r-ball at x as $W_r(x) = \{z \in K : ||z - x||_x \leq r\}$. Then the 1-ball $W_1(x)$ is called a Dikin ellipsoid that reflects the curvature of R at x in the length of its axes. A nice fact about the Dikin ellipsoid here is that it will always lie inside the set K. For this reason it is particularly relevant to interior point methods. For our purpose, the Dikin ellipsoid tells us in which directions we have a lot of space to move and we will adapt our sampling strategy according to its induced local geometry. Specifically, if $\{\lambda_1, \ldots, \lambda_n\}$ and $\{e_1, \ldots, e_n\}$ are eigenvalues and eigenvectors of $\nabla^2 R(x_t)$, then we let $y_t = x_t + \epsilon_t \lambda_{i_t}^{-1/2} e_{i_t}$ and estimate the slope as $\tilde{l}_t = n(l_t^\top y_t)\epsilon_t \lambda_{i_t}^{1/2}$. The key to bounding the regret will be that the divergence is approximately $D_{R^*}(0, \eta\tilde{l}_t) \approx \eta^2 \tilde{l}_t (\nabla^2 R(x_t))^{-1} \tilde{l}_t$. Here, the inverse Hessian will kill the largeness of $\tilde{l}_t$ in exactly the right directions, leaving us with terms on the order of η^2.

$$\begin{aligned} D_{R^*}(0, \eta\tilde{l}_t) &\approx \eta^2 \tilde{l}_t (\nabla^2 R(x_t))^{-1} \tilde{l}_t & (15)\\ &= \eta^2 n^2 (l_t^\top y_t)^2 \lambda_{i_t} e_{i_t}^\top (\nabla^2 R(x_t))^{-1} e_{i_t} & (16)\\ &= \eta^2 n^2 (l_t^\top y_t)^2. & (17) \end{aligned}$$

This latter property relies on our sampling strategy along the eigendirections and would not hold if we were to mix eigendirections. In brief, the regulariser gives us a geometry so that we can sample and through the Hessian gives us control on the regret.

The full details for this lecture can be found Abernethy et al. [1]. Of specific relevance are Lemma 4.1 as well as Theorem 3.1. A proof sketch for Theorem 3.1 is given in Section 6 and the complete argument in Section 8.

References

[1] Jacob Duncan Abernethy, Elad Hazan, and Alexander Rakhlin. An efficient algorithm for bandit linear optimization. Technical Report UCB/EECS-2008-18, EECS Department, University of California, Berkeley, Feb 2008.