6.883: Online Methods in Machine Learning

Alexander Rakhlin

LECTURE 11

0.1 Continued: Binary prediction with indicator loss and no side information

Recall that we are considering here the simplest possible online scenario

For $t = 1, \dots, n$
  Predict $\widehat{y}_t \in \{\pm 1\}$
  Observe outcome $y_t \in \{\pm 1\}$

and we employ a randomized strategy defined by the choice of the mean $q_t(y_1, \dots, y_{t-1}) \in [-1, 1]$ of the distribution for $\widehat{y}_t$.

The optimization of

$$\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{\widehat{y}_t \neq y_t\} \tag{1}$$

cannot be done for all sequences, and so we choose a function $\phi : \{\pm 1\}^n \to \mathbb{R}$ which tells us which sequences we care about. The goal, once again, is to find an algorithm such that

$$\forall y_1, \dots, y_n, \quad \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{\widehat{y}_t \neq y_t\}\right] \leq \phi(y_1, \dots, y_n). \tag{2}$$

We already presented a closed-form strategy for this in Lecture 1, but it appeared out of thin air. We now present a derivation of this strategy. We also write the optimization problem in a *minimax form* that will be important for the rest of the course.

First, we write (2) as a telescoping sum

$$\forall y_1, \dots, y_n, \quad \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{\widehat{y}_t \neq y_t\}\right] \leq \sum_{t=1}^{n} \mathbf{Rel}_{t-1}(y_{1:t-1}) - \mathbf{Rel}_t(y_{1:t}) \tag{3}$$

with some *yet to be determined* functions $\mathbf{Rel}_t$, satisfying

$$\mathbf{Rel}_n = -\phi \quad \text{and} \quad \mathbf{Rel}_0(\varnothing) \leq 0. \tag{4}$$

At this point we have not lost any generality by going from (2) to (3), since all the functions telescope.

Let us drop the subscript on $\mathbf{Rel}$, e.g. by defining $\mathbf{Rel}$ to be a function $\cup_{t=1}^{n}\{\pm 1\}^t \to \mathbb{R}$.

Suppose $y_1, \dots, y_{n-1}$ have been revealed, and we are about to predict the last bit. The goal (3) can be phrased as: there *exists* a way to choose a randomized strategy q_n such that *for any* y_n (3) holds. Translating these quantifiers into minima and maxima (something we will do all the time from now on), leads to

$$\min_{q_n} \max_{y_n} \left\{ \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{\widehat{y}_t \neq y_t\}\right] - \sum_{t=1}^{n} \left[\mathbf{Rel}(y_{1:t-1}) - \mathbf{Rel}(y_{1:t})\right] \right\} \leq 0 \tag{5}$$

Once again, the expectations are with respect to the randomizations of the algorithm. Now observe that the terms $1:n-1$ are not involved in the minimization/maximization, and the last-step optimization problem is

$$\min_{q_n} \max_{y_n} \left\{ \mathbb{E}\left[\frac{1}{n} \mathbf{1}\{\widehat{y}_n \neq y_n\} \right] + \mathbf{Rel}(y_{1:n}) \right\} \tag{6}$$

Recall that we are free to define $\mathbf{Rel}(y_{1:n-1})$. We could simply define it as the value of the above objective function. More generally, $\mathbf{Rel}(y_{1:n-1})$ will be an upper bound on it:

$$\min_{q_n} \max_{y_n} \left\{ \mathbb{E}\left[\frac{1}{n} \mathbf{1}\{\widehat{y}_n \neq y_n\} \right] + \mathbf{Rel}(y_{1:n}) \right\} \leq \mathbf{Rel}(y_{1:n-1}). \tag{7}$$

The inequality (7) *is key and will appear for the rest of the course, so it's good to understand how it came about.*

Since $\widehat{y}_n, y_n \in \{\pm 1\}$, we write

$$\mathbb{E}\left[\mathbf{1}\{\widehat{y}_n \neq y_n\} \right] = \mathbb{E}\left[\frac{1}{2}\left(1 - \widehat{y}_n y_n\right) \right] = \frac{1}{2} - \frac{1}{2} q_n y_n \tag{8}$$

since q_n is precisely the mean of $\widehat{y}_n$. Then

$$\min_{q_n} \max_{y_n} \left\{ \mathbb{E}\left[\frac{1}{n} \mathbf{1}\{\widehat{y}_n \neq y_n\} \right] + \mathbf{Rel}(y_{1:n}) \right\} \tag{9}$$

$$= \min_{q_n} \max \left\{ -\frac{1}{2n} q_n + \mathbf{Rel}(y_{1:n-1}, +1), \quad \frac{1}{2n} q_n + \mathbf{Rel}(y_{1:n-1}, -1) \right\} + \frac{1}{2n} \tag{10}$$

The minimum over $q_n \in [-1, 1]$ of two linear functions with opposite slopes will occur when they are equal, given that the minimum belongs to the interval $[-1, 1]$. The latter requirement will be guaranteed by the smoothness assumption on ϕ once we get a closed form for **Rel**. We have

$$q_n(y_{1:n-1}) = n(\mathbf{Rel}(y_{1:n-1}, +1) - \mathbf{Rel}(y_{1:n-1}, -1)).$$

Plugging this solution into the minimax expression,

$$\min_{q_n} \max_{y_n} \left\{ \mathbb{E}\left[\frac{1}{n} \mathbf{1}\{\widehat{y}_n \neq y_n\} \right] + \mathbf{Rel}(y_1, \ldots, y_n) \right\} \tag{11}$$

$$= \frac{1}{2}(\mathbf{Rel}(y_{1:n-1}, +1) + \mathbf{Rel}(y_{1:n-1,-1})) + \frac{1}{2n} \tag{12}$$

$$= \mathbb{E}_{\epsilon_n} \mathbf{Rel}(y_{1:n-1}, \epsilon_n) + \frac{1}{2n} \tag{13}$$

$$= -\mathbb{E}_{\epsilon_n} \phi(y_{1:n-1}, \epsilon_n) + \frac{1}{2n} \tag{14}$$

$$\triangleq \mathbf{Rel}(y_{1:n-1}) \tag{15}$$

where the last step is now a definition. We have verified (7) and came up with a definition to ensure it; smoothness of ϕ also played a role in certifying a solution for the minimum. The requirement (5) now reads

$$\mathbb{E}\left[\frac{1}{n} \sum_{t=1}^{n-1} \mathbf{1}\{\widehat{y}_t \neq y_t\} \right] - \sum_{t=1}^{n-1} \left[\mathbf{Rel}(y_{1:t-1}) - \mathbf{Rel}(y_{1:t}) \right] \leq 0 \tag{16}$$

and we may now proceed to analyse step $n-1$. This is certainly the same step as in Lecture 1, but we have explicitly derived the solution q_n for the last step. Applying the operators $\min_{q_{n-1}} \max_{y_{n-1}}$ to (16) gives

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n-2}\mathbf{1}\{\widehat{y}_t \neq y_t\}\right] - \sum_{t=1}^{n-2}\left[\mathbf{Rel}(y_{1:t-1}) - \mathbf{Rel}(y_{1:t})\right] \leq 0. \tag{17}$$

and so on. We see that for step t

$$\mathbf{Rel}(y_{1:t}) = -\mathbb{E}_{\epsilon_{t+1:n}}\phi(y_{1:t}, \epsilon_{t+1:n}) + \frac{n-t}{2n} \tag{18}$$

and the requirement that $\mathbf{Rel}(\varnothing) \leq 0$ is

$$\mathbf{Rel}(\varnothing) = -\mathbb{E}\phi(\epsilon_{1:n}) + \frac{1}{2} \leq 0,$$

which is precisely the assumption on ϕ.

Remark 1. *We note that the requirement* $\mathbf{Rel}_0 \leq 0$ *in* (4) *may be dropped. In this case, if we define* $\mathbf{Rel}_n = -\phi$*, the final bound in* (3) *will be* $\phi + \mathbf{Rel}_0$*, a constant shift of* ϕ*. This small variation will be convenient later when the potential is known only up to a constant shift. In particular, if we use* $d_H(\boldsymbol{y}, F)$ *as the potential, we simply take* $\mathbf{Rel}_n = -d_H(\boldsymbol{y}, F)$ *and then observe that* $\mathbf{Rel}_0 = C_n$*, yielding the final bound of* $d_H(\boldsymbol{y}, F) + C_n$*.*

We have achieved several goals here: first, we proved that $\mathbb{E}\phi \geq \frac{1}{2}$ is a sufficient (and necessary) condition; second, we derived the strategy by solving a minimax expression from inside out ($t = n$ to $t = 1$); third, we phrased the problem of Lecture 1 in the language of **Rel**, which we shall later call *relaxations*. The function **Rel** may also be called a *potential function*, and the key inequality (7) may be interpreted as: on a given round, the change in the potential is at least the size of the expected mistake.

0.2 Example: Prediction on Graphs

Consider the following setting. There is a known connected graph $G = (V, E)$, with $n = |V|$. At each time step, we are pointed to a vertex and required to predict its label. We do not come back to the same vertex twice, and so the time horizon n is equal to the number of vertices. Suppose that the order in which the vertices appear is known to us (below, we will argue that this knowledge is not needed).

We would like to make a small number of mistakes, and the prior information we would like to use is that *neighbors tend to have same labels*. This feature of networks, called homophily, may be encoded in a probabilistic manner via a generative process, but here we are encoding the assumption in the function ϕ. Given a sequence $y_1, \ldots, y_n$, how can we tell if it adheres to the homophily assumption? Well, we can measure how often neighbors disagree under this labeling. Let $y_{(v)}$ stand for the label given by $\boldsymbol{y} = (y_1, \ldots, y_n)$ to the vertex $v \in V$. Then the number of times neighbors disagree is simply

$$\frac{1}{4}\sum_{(u,v)\in E}(y_{(v)} - y_{(u)})^2$$

This quadratic form may be written as

$$\sum_{(u,v)\in E}(y_{(v)} - y_{(u)})^2 = \boldsymbol{y}^\top L \boldsymbol{y},$$

where $L = D - A$ is the graph Laplacian, defined as a difference of a degree matrix D (diagonal elements corresponding to degrees) and an adjacency matrix A. Each labeling $\boldsymbol{y}$ defines a *cut* – all the edges with disagreeing labels at the ends. We will thus refer to the above quadratic form as the *size of the cut* induced by $\boldsymbol{y}$.

More generally, if the graph is weighted, with a weight $w_{(u,v)}$ on the edge (u, v),

$$\sum_{(u,v)\in E} w_{(u,v)}(y_{(v)} - y_{(u)})^2 = \boldsymbol{y}^\top L\boldsymbol{y}.$$

We now have a way to decide which sequences are more important to us, given the homophily assumption. Here are two approaches that come to mind (think about finding others!)

- first is to define a "nice" set

$$F_K = \left\{f \in \{\pm 1\}^n : f^\top L f \le K\right\} \tag{19}$$

  for some parameter K and take

$$\phi(\boldsymbol{y}) = d_H(\boldsymbol{y}, F) + C_n, \quad C_n = \frac{1}{2n}\mathbb{E}\max_{f\in F}\langle f, \boldsymbol{\epsilon}\rangle \tag{20}$$

- second approach is to define $\phi(\boldsymbol{y})$ directly as some function of the value $\boldsymbol{y}^\top L\boldsymbol{y}$, say

$$\phi(\boldsymbol{y}) = a_n \boldsymbol{y}^\top L\boldsymbol{y} \tag{21}$$

  for n-dependent a_n

The first approach satisfies all the conditions required for the existence of a strategy, as it fits squarely in the development of the previous section. By construction, the function is smooth and its expected value is above 1/2.

The second approach, on the other hand, is not in the form of a normalized Hamming distance. In fact, as we now show, the function in (21) cannot be made to satisfy both conditions. What happens to the size of the cut as we change one label? The cut can change by at most the degree of the vertex. So, assuming that G has degree at most d, we need to take $a_n \le \frac{1}{dn}$. On the other hand, the expected size of a cut for a random labeling is $|E|/2$ (prove this), and thus

$$a_n \frac{|E|}{2} \ge \frac{1}{2}.$$

Yet, there are methods in the literature that guarantee that the number of mistakes is bounded by (a scaling of) cut size. There is no contradiction here, as Cover's equivalence was only holding under the assumption that the function is smooth. Trivially, a function $\phi(\boldsymbol{y}) = \mathbf{1}\{\boldsymbol{y} \ne \mathbf{1}\}$ is one such nonsmooth function that admits a prediction strategy with a bound (2) (predict 1 on every round).

To get a sense of how the two approaches may differ on a particular example, consider a star graph, made by taking a center node and connecting it to $n-1$ other vertices. Consider the labeling $\boldsymbol{y}$ that assigns -1 to the center and $+1$ to the other nodes. The size of the cut is $n-1$, and the algorithm is allowed to make that many mistakes on the sequence. However, for any $K < n-1$, the set F_K contains two labelings (all -1 and all $+1$), and the sequence $\boldsymbol{y}$ described above is 1-Hamming distance away from F_K. Hence, the average number of mistakes guaranteed for the algorithm is $\frac{1}{n} + \frac{c}{\sqrt{n}}$ (c can be computed as Rademacher average of two point class).