6.883: Online Methods in Machine Learning

Alexander Rakhlin

LECTURES 15 AND 16

1. THE EXPERTS SETTING. EXPONENTIAL WEIGHTS

All the algorithms presented so far hallucinate the future values as random draws and then perform two evaluations of ϕ. In many situations, an easier method is available—one that does not require drawing the random variables. In fact, most of the methods encountered in the online learning literature can be seen as doing precisely this—getting rid of the random variables for "future" rounds.

One of the most famous scenarios in online learning is that of *prediction with expert advice*. There are several roughly equivalent formulations, but once you've seen one, you'll be able to modify the proof for the other.

Consider the situation where on each round $t = 1, \dots, n$, we observe advice of N experts. Suppose the advice comes in the form of a vector $x_t \in [-1, 1]^N$, and we think of $x_t(i)$ as, say, the buy/sell advice by expert i. We treat x_t as side information for making our own decision. After seeing the advice, we decide on a mixed strategy $\widehat{y}_t \in \Delta(N)$ (a distribution over the N experts) and make a prediction $\langle \widehat{y}_t, x_t \rangle \in [-1, 1]$ by mixing the opinions according to $\widehat{y}_t$. The outcome $y_t \in \{\pm 1\}$ is then revealed.

Once we have the mean $\langle \widehat{y}_t, x_t \rangle$ for the mixed strategy, we may either draw the actual binary-valued prediction from this distribution, or we may simply think of

$$|y_t - \langle \widehat{y}_t, x_t \rangle| = 1 - \langle \widehat{y}_t, y_t x_t \rangle$$

as the expected indicator loss of our strategy (see the collaborative filtering example). What is different here from previous lectures is that our decision variable $\widehat{y}_t$ is not a real number, but a distribution.

The goal of the learner is to incur small average loss

$$\frac{1}{n} \sum_{t=1}^{n} |y_t - \langle \widehat{y}_t, x_t \rangle|. \tag{1}$$

As it turns out, a simple algorithm allows the learner to keep this average loss *not much worse than the loss of the best expert*, without knowing who the best is until the end. In particular, we prove that

Lemma 1. *There is an algorithm (in fact, several distinct methods) that guarantees*

$$\frac{1}{n} \sum_{t=1}^{n} |y_t - \langle \widehat{y}_t, x_t \rangle| \le \min_{j \in [N]} \frac{1}{n} \sum_{t=1}^{n} |y_t - x_t(j)| + c \sqrt{\frac{\log N}{n}} \tag{2}$$

for any sequence $(x_1, y_1), \dots, (x_n, y_n)$. As an example, this bound (with $c = \sqrt{8}$) is attained by the exponential weights algorithm

$$\widehat{y}_t(j) \triangleq \frac{e^{-\eta \sum_{s=1}^{t-1} |y_s - x_s(j)|}}{\sum_{j=1}^{N} e^{-\eta \sum_{s=1}^{t-1} |y_s - x_s(j)|}} \tag{3}$$

with a step size $\eta = \sqrt{\frac{\log N}{2n}}$.

We will present two very similar proofs. After proving the Lemma, we will re-do the proof in the slightly simpler transductive setting through the lens of Cover's statement. It's instructive to look at both proofs and see the few small differences.

Both proofs will utilize the following inequalities. First is the *soft-max* bound. Choose a parameter $\eta > 0$ and let $A_1, \ldots, A_N$ be real numbers. We then have

$$\max_{j\in[N]} A_j = \frac{1}{\eta}\max_j \eta A_j = \frac{1}{\eta}\log\exp\max_j \eta A_j = \frac{1}{\eta}\log\max_j \exp\{\eta A_j\} \le \frac{1}{\eta}\log\sum_{j=1}^{N}\exp\{\eta A_j\}.$$

There is only one inequality between the maximum over j and the "soft-max" function. Suppose all A_j are equal. Then the right-hand-side is larger than the left-hand-side by an additive $\eta^{-1}\log N$ factor (verify this!). As η increases, the gap between the two sides vanishes. Same can be argued for the case when the values are not equal. In fact, the last upper bound is an equality if we are allowed to choose η.

The second inequality we use is

$$\frac{1}{2}(e^x + e^{-x}) \le e^{x^2/2}$$

which you can prove via Taylor expansions. The inequality implies

$$\mathbb{E}e^{\lambda\epsilon} \le e^{\lambda^2/2} \tag{4}$$

for the Rademacher random variable ϵ and a contant $\lambda \in \mathbb{R}$. The same bound holds for *any* zero-mean random variable Z with values in $[-1, 1]$:

$$\mathbb{E}e^{\lambda Z} \le e^{\lambda^2/2}, \tag{5}$$

and it is immediate that the bound becomes $e^{b^2\lambda^2/2}$ for $[-b, b]$-valued Z.

1.1 Proof of Lemma 1

Thanks to the identity

$$|a - b| = 1 - ab$$

that holds for $a \in \{\pm 1\}$ and $b \in [-1, 1]$, we may rewrite the difference between the loss of the algorithm and the benchmark in (2) as

$$\frac{1}{n}\sum_{t=1}^{n} -y_t\langle\widehat{y}_t, x_t\rangle - \min_{j\in[N]}\frac{1}{n}\sum_{t=1}^{n} -y_t x_t(j). \tag{6}$$

Let us omit the fraction $\frac{1}{n}$ and bring it back at the very end. When comparing to the proof in the next section, just insert this fraction throughout.

Consider the last step $t = n$. In the first sum, all the terms except the last one are fixed, and so we need to solve

$$\min_{\widehat{y}_n\in\Delta(N)}\max_{y_n\in\{\pm1\}}\{-y_n\langle\widehat{y}_n, x_n\rangle + \mathbf{Rel}(x_{1:n}, y_{1:n})\} \tag{7}$$

with

$$\mathbf{Rel}(x_{1:n}, y_{1:n}) \geq -\min_{j\in[N]} \sum_{t=1}^{n} -y_t x_t(j). \tag{8}$$

Here $\Delta(N)$ is the probability simplex on N experts. We could choose **Rel** to be equal to the right-hand side in (8). However, for computational purposes, we slightly modify this function. Instead of max we shall work with "soft-max". That is, take

$$\mathbf{Rel}(x_{1:n}, y_{1:n}) = \frac{1}{\eta} \log \sum_{j=1}^{N} \exp\left\{\eta \sum_{t=1}^{n} y_t x_t(j)\right\}$$

for some η, to be determined.

The algorithm (3), in the form (6), can be written as

$$\widehat{y}_t(j) \propto e^{\eta \sum_{s=1}^{t-1} y_s x_s(j)} \tag{9}$$

while the loss on round t is (trivially)

$$-y_t \langle \widehat{y}_t, x_t \rangle = -\mathbb{E}_{j\sim\widehat{y}_t}\left[y_t x_t(j)\right] = \frac{1}{\eta} \log \exp\left\{-\eta \mathbb{E}_{j\sim\widehat{y}_t}\left[y_t x_t(j)\right]\right\}.$$

The key observation now is that

$$\sum_{j=1}^{N} \exp\left\{\eta \sum_{t=1}^{n} y_t x_t(j)\right\} = \sum_{j=1}^{N} \exp\left\{\eta y_n x_n(j)\right\} \exp\left\{\eta \sum_{t=1}^{n-1} y_t x_t(j)\right\} \tag{10}$$

$$= \mathbb{E}_{j\sim\widehat{y}_n}\left[\exp\left\{\eta y_n x_n(j)\right\}\right] \times \sum_{j=1}^{N} \exp\left\{\eta \sum_{t=1}^{n-1} y_t x_t(j)\right\} \tag{11}$$

because $\mathbb{E}_{j\sim\widehat{y}_n}\left[A(j)\right] = \sum_{j=1}^{N} \widehat{y}_n(j) A(j)$ with

$$\widehat{y}_n(j) = \frac{\exp\left\{\eta \sum_{t=1}^{n-1} y_t x_t(j)\right\}}{\sum_{j=1}^{N} \exp\left\{\eta \sum_{t=1}^{n-1} y_t x_t(j)\right\}}. \tag{12}$$

Putting everything together, (7) is upper bounded by the particular choice of the infimum strategy $\widehat{y}_t$ as

$$(7) \leq \max_{y_n\in\{\pm1\}} \left\{\langle \widehat{y}_n, -y_n x_n\rangle + \mathbf{Rel}(x_{1:n}, y_{1:n})\right\} \tag{13}$$

$$= \max_{y_n\in\{\pm1\}} \left\{\frac{1}{\eta} \log \exp\left\{-\eta \mathbb{E}_{j\sim\widehat{y}_n}\left[y_n x_n(j)\right]\right\} + \frac{1}{\eta} \log \mathbb{E}_{j\sim\widehat{y}_n}\left[\exp\left\{\eta y_n x_n(j)\right\}\right]\right\} \tag{14}$$

$$+ \frac{1}{\eta} \log \sum_{j=1}^{N} \exp\left\{\eta \sum_{t=1}^{n-1} y_t x_t(j)\right\} \tag{15}$$

We now focus on the two terms in (14):

$$\frac{1}{\eta} \log \exp\left\{-\eta \mathbb{E}_{j\sim\widehat{y}_n}\left[y_n x_n(j)\right]\right\} + \frac{1}{\eta} \log \mathbb{E}_{j\sim\widehat{y}_n}\left[\exp\left\{\eta y_n x_n(j)\right\}\right] \tag{16}$$

$$= \frac{1}{\eta} \log \mathbb{E}_{j\sim\widehat{y}_n}\left[\exp\left\{\eta(y_n x_n(j) - \mathbb{E}_{j\sim\widehat{y}_n}\left[y_n x_n(j)\right])\right\}\right] \tag{17}$$

$$\leq \frac{1}{\eta} \frac{(2\eta)^2}{2} = 2\eta \tag{18}$$

by (5). Note that the range of the zero-mean random variable is $[-2, 2]$, and so an additional factor of 2^2 appears from the application of (5). Observe that this last step of peeling off the zero-mean term makes the expression *independent of* y_n *and* x_n! In particular, *it does not matter whether the sequence of* x*'s is generated i.i.d. or in an arbitrary manner.*

Unlike the Cover's approach of solving the max over the two alternatives, we presented a particular $\widehat{y}_t$ that allows (through an upper bound) to make the choice y_n irrelevant. While the two approaches give slightly different algorithms, the upper bounds they enjoy are the same.

Now, we simply define

$$\mathbf{Rel}(x_{1:n-1}, y_{1:n-1}) = \frac{1}{\eta} \log \sum_{j=1}^{N} \exp\left\{ \eta \sum_{t=1}^{n-1} y_t x_t(j) \right\} + 2\eta \tag{19}$$

and

$$\mathbf{Rel}(x_{1:t}, y_{1:t}) = \frac{1}{\eta} \log \sum_{j=1}^{N} \exp\left\{ \eta \sum_{s=1}^{t} y_s x_s(j) \right\} + 2(n-t)\eta. \tag{20}$$

Of course,

$$\mathbf{Rel}(\varnothing) = \frac{1}{\eta} \log N + 2n\eta \le 2\sqrt{2n \log N}. \tag{21}$$

by choosing $\eta = \sqrt{\frac{\log N}{2n}}$. Now, since we initially divided throughout by $1/n$, the bound of the lemma is $\sqrt{\frac{8 \log N}{n}}$.

1.2 (slightly easier) transductive setting through the lens of Cover's algorithm

Consider the simplified setting where expert advices $x_1, \dots, x_n \in \{\pm 1\}^N$ are fixed and known a priori, and let $\mathcal{F} = \{x \mapsto x_j : j \in [N]\}$ be the set of N functions that simply output a coordinate of x. As discussed earlier, $\mathcal{F}$ induces a subset $F \subseteq \{\pm 1\}^n$ of finite cardinality N,

$$F = \{\boldsymbol{y} : \boldsymbol{y} = (x_1(j), \dots, x_n(j)), j \in [N]\}$$

and

$$\phi(\boldsymbol{y}) = d_H(\boldsymbol{y}, F) + C_n = \min_{f \in F} \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\{f_t \neq y_t\} + C_n \tag{22}$$

for some appropriate C_n which we will define later.

In this section, we will directly solve for the real-valued prediction $q_t \in [-1, 1]$, which can be viewed as the mean of the mixed strategy for predicting $\widehat{y}_t \in \{\pm 1\}$. This is slightly different from what was described earlier, where the prediction is calculated by mixing the advice as $\langle \widehat{y}_t, x_t \rangle$. We will solve the latter directly in the next section.

Recall that the choice of relaxation **Rel** defines the algorithm. In this section we give the derivation using the very basic technique that goes back to Lecture 1. The algorithm that arises from Cover's lemma is not exponential weights, but it gives the same guarantee on performance as the exponential weights method.

Let us take $\mathbf{Rel}(y_{1:n})$ as any upper bound on the benchmark term

$$-\min_{f\in F}\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}\{f_t \neq y_t\} = -\frac{1}{2}+\frac{1}{2n}\max_{f\in F}\langle f,\boldsymbol{y}\rangle\,. \tag{23}$$

We will use the *soft-max upper bound* (verify that it holds):

$$\mathbf{Rel}(y_{1:n}) \triangleq -\frac{1}{2}+\frac{1}{2n}\frac{1}{\eta}\log\sum_{f\in\mathcal{F}}\exp\{\eta\langle f,\boldsymbol{y}\rangle\} \tag{24}$$

Check that this function does not change by more than $1/n$ when flipping one bit. Now, as before,

$$\min_{q_n}\max_{y_n}\left\{\mathbb{E}\left[\frac{1}{n}\mathbf{1}\{\widehat{y}_n \neq y_n\}\right]+\mathbf{Rel}(y_{1:n})\right\} = \mathbb{E}_{\epsilon_n}\mathbf{Rel}(y_{1:n-1},\epsilon_n)+\frac{1}{2n} \tag{25}$$

By Jensen's inequality ($\mathbb{E}\log \le \log\mathbb{E}$),

$$\mathbb{E}_{\epsilon_n}\mathbf{Rel}(y_{1:n-1},\epsilon_n) \le -\frac{1}{2}+\frac{1}{2n}\frac{1}{\eta}\log\sum_{f\in\mathcal{F}}\mathbb{E}_{\epsilon_n}\exp\{\eta\langle f,\tilde{\boldsymbol{y}}\rangle\} \tag{26}$$

with $\tilde{\boldsymbol{y}} = (y_{1:n-1},\epsilon_n)$. The only randomness in the above expression is the ϵ_n on the last coordinate of $\tilde{\boldsymbol{y}}$. Let us abuse the notation and write $\langle f,\tilde{\boldsymbol{y}}\rangle = \langle f_{1:n-1},y_{1:n-1}\rangle + \epsilon_n f_n$. *Our aim is to get rid of* ϵ_n. If we succeed, we do not have to draw the random coin flips for random playout at the intermediate steps.

By (4),

$$\mathbb{E}_{\epsilon_n}\exp\{\eta\langle f,\tilde{\boldsymbol{y}}\rangle\} \le \exp\{\eta\langle f_{1:n-1},y_{1:n-1}\rangle\}\times\exp\{\eta^2/2\} \tag{27}$$

and, therefore, (26) is upper bounded by

$$\frac{1}{2n}\frac{1}{\eta}\log\sum_{f\in F}\exp\{\eta\langle f_{1:n-1},y_{1:n-1}\rangle\}+\frac{\eta}{4n}-\frac{1}{2} \tag{28}$$

In view of (25), we can now define,

$$\mathbf{Rel}(y_{1:n-1}) = \frac{1}{2n\eta}\log\sum_{f\in F}\exp\{\eta\langle f_{1:n-1},y_{1:n-1}\rangle\}+\frac{\eta}{4n}-\frac{n-1}{2n}. \tag{29}$$

That's it! There is no ϵ_n in the relaxation at time $n-1$. We "peeled it off". One can check that at the intermediate step t,

$$\mathbf{Rel}(y_{1:t}) = \frac{1}{2n\eta}\log\sum_{f\in F}\exp\{\eta\langle f_{1:t},y_{1:t}\rangle\}+\frac{(n-t)\eta}{4n}-\frac{t}{2n} \tag{30}$$

with $\langle f_{1:t},y_{1:t}\rangle$ being defined as $\sum_{s=1}^{t} f_s y_s$. We also see that

$$\mathbf{Rel}(\varnothing) = \frac{1}{2n\eta}\log\sum_{f\in F}\exp\{0\}+\frac{\eta}{4} = \frac{1}{2n\eta}\log N+\frac{\eta}{4} = \sqrt{\frac{\log N}{2n}} \tag{31}$$

by choosing $\eta = \sqrt{\frac{2\log N}{n}}$. This is a non-algorithmic derivation, and the algorithm is given in Lecture 1. We leave it as a homework exercise to write it explicitly. (hint: it does not become exponential weights). We also note that the difference in the constant c comes from scaling of the indicator loss vs the absolute value loss.

1.3 Discussion

The two proofs are essentially the same. Both start by relaxing the max to a soft-max, and taking this as $\mathbf{Rel}_n$. Then, the second approach explicitly solves for the optimal real-valued prediction (the mean of the mixed strategy), while the first approach guesses a (potentially suboptimal) strategy of exponential weights. Once the strategy for the infimum is plugged in, one obtains an expression with a zero-mean random variable. This zero-mean variable is eliminated using a probabilistic inequality (Eq. (17) and (27), respectively).

To reiterate, the salient features of the proofs are: (1) passing to a relaxation for $\mathbf{Rel}_n$, (2) solving for the best strategy or guessing a near-best strategy, and (3) using probabilistic inequalities to remove the random variable that arises from plugging in the strategy. *These steps can be taken as a rough prescription for the development of online methods.* We will illustrate the steps again in the subsequent lectures.

The next note is on the nature of the sequence $x_1, \ldots, x_n$. Essentially, both approaches make it irrelevant how the x_t's are generated. That is not to say that the method does not take the side information into account (of course it does – through the losses of experts). Rather, the point is that we can successfully deal with adversarially generated x's, a strength of the experts approach.

Another strength of the experts bound is its mild (logarithmic) dependence on the number of experts. One may take a large number of experts and still have an average error being $o(1)$ from the average error of the best expert.

Finally, we remark that the experts approach can be seen as a "union bound" or an aggregation procedure. Suppose one has N algorithms making predictions. Then one can predict as well as any of these algorithms by paying $O\left(\sqrt{\frac{\log N}{n}}\right)$. Such a black box technique is very useful (see the version of "linearized experts" below for the general black box statement). For instance, suppose one does not know how to choose a parameter $\theta \in [0,1]$ of the algorithm optimally. One can then run (at least in principle) $N = 1/\epsilon$ algorithms corresponding to an ϵ-discretization of the parameter. If the output is in some sense Lipschitz with respect to the parameter choice, on can claim that the resulting aggregating procedure does as well as the best choice, plus an ϵ-precision term, plus an $O\left(\sqrt{\frac{\log(1/\epsilon)}{n}}\right)$ penalty.

1.4 Linearized Experts

Recall that in the experts setting introduced in the beginning of the lecture, we observe predictions $x_t \in [-1,1]^N$ of the experts, choose a distribution $\widehat{y}_t \in \Delta(N)$ for the weighted vote, and then observe the outcome $y_t \in \{\pm 1\}$. Observe that the exponential weights algorithm at time t does not use x_t to calculate the distribution over experts. Hence, we may think of a setting where we choose a distribution $\widehat{y}_t \in \Delta(N)$ and then observe both predictions x_t and the true outcome y_t. Rather than mixing the advice of the experts to produce our own, we may instead choose the expert at random from the distribution $\widehat{y}_t$ and go with her advice. Then the expected cost for the period t is

$$\langle \widehat{y}_t, z_t \rangle$$

where $z_t \in [-1,1]^N$ is the vector of losses for each expert: $z_t(j) = |y_t - x_t(j)|$. In fact, the loss function does not matter anymore, and it does not matter that data comes in the form (x_t, y_t) pairs. Instead, we may just think of each expert incurring some cost, we are

choosing an expert at random and incur the same cost as that expert. In expectation, we pay $\langle \widehat{y}_t, z_t \rangle$.

Let us state the protocol explicitly:

For $t = 1, \dots, n$
  Predict $\widehat{y}_t \in \Delta(N)$
  Observe costs $z_t \in [-1, 1]^N$

Alternatively, we may choose the random expert j according to $\widehat{y}_t$ and pay $z_t(j)$.

The goal here is to have small expected cost, relative to the cost of the best expert:

$$\frac{1}{n} \sum_{t=1}^{n} \langle \widehat{y}_t, z_t \rangle \leq \min_{j \in [N]} \frac{1}{n} \sum_{t=1}^{n} \langle \boldsymbol{e}_j, z_t \rangle + c \sqrt{\frac{\log N}{n}} \tag{32}$$

for any sequence $z_1, \dots, z_n$ of costs. The cost z_t may be chosen even with the knowledge of our decision $\widehat{y}_t$.

Let us quickly prove that the exponential weights algorithm

$$\widehat{y}_t(j) \propto \exp\left\{ -\eta \sum_{s=1}^{t-1} z_s(j) \right\}$$

achieves the above guarantee. We write the last step of the problem (removing the $1/n$ normalization term) as

$$\min_{\widehat{y}_n \in \Delta(N)} \max_{z_n \in [-1,1]^N} \{ \langle \widehat{y}_n, z_n \rangle + \mathbf{Rel}(z_{1:n}) \} \tag{33}$$

with

$$-\min_{j \in [N]} \sum_{t=1}^{n} \langle \boldsymbol{e}_j, z_t \rangle = \max_{j \in [N]} -\sum_{t=1}^{n} z_t(j) \leq \frac{1}{\eta} \log \sum_{j=1}^{N} \exp\left\{ -\eta \sum_{t=1}^{n} z_t(j) \right\} \triangleq \mathbf{Rel}(z_{1:n}). \tag{34}$$

The rest of the proof of (32) is essentially identical to that of the proof of Lemma 1.

References