10-716: Advanced Machine Learning **Spring 2019**

Lecture 21: April 16

Lecturer: Pradeep Ravikumar *Scribes: Boxiang Lyu*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

21.1 Example

Recall the example that we discussed in the previous lecture, where we assume that there is an expert that always makes the right prediction. However, the assumption is fairly strong, and sometimes even the best expert can make mistakes.

Consider another setting where, after round t, set $w_i^t = \beta w_i^{t-1}$ if $f_{i,t} \neq y_t$, where $\beta \in (0,1)$ is an arbitrary parameter. The forecast we make at round t is the weighted average of the experts' advice with weights $w_1^t, \ldots, w_N^t$.

Claim 21.1 *The number of rounds in which the forecaster makes a mistake is* $m \leq \frac{m^* \log \frac{1}{\beta} + \log N}{\log \frac{2}{1+\beta}}$, *where* m^* *is the number of mistakes the best expert has made up to this point.*

Proof: Similar to the proof we did in the last lecture, let W_m be the sum of the weights of all experts after the forecaster has made m mistakes. Initially, $m = 0$, $w_{i,0} = 1$ for $i \in \{1, 2, \ldots N\}$, and $W_0 = N$. We then have that

$$\begin{aligned} W_m &= \sum_{i=1}^{N} w_{i,m} \\ &\leq \beta \frac{W_{m-1}}{2} + \frac{W_{m-1}}{2} = \left(\frac{1+\beta}{2}\right) W_{m-1} \\ &\leq \left(\frac{1+\beta}{2}\right)^m W_0 = \left(\frac{1+\beta}{2}\right)^m N \end{aligned}$$

Let m^* be the number of mistakes made by the best expert when the forecaster has made m mistakes. The weight of the expert is then β^{m^*} by the algorithm, and we have

$$\beta^{m^*} \leq W_m \leq \left(\frac{1+\beta}{2}\right)^m N$$

Solving the inequality gives

$$m \leq \frac{m^* \log \frac{1}{\beta} + \log N}{\log \frac{2}{1+\beta}}$$

■

21.2 Weighted Average Forecaster

The weighted average forecaster predicts at time t according to

$$\hat{p}_t = \frac{\sum_{i=1}^N w_{i,t-1} f_{i,t}}{\sum_{i=1}^N w_{i,t-1}}$$

, which can be viewed as a convex combination of expert advice $\{f_{1,t}, f_{2,t}, \ldots, f_{n,t}\}$. Assume that the decision space $\mathcal{D}$ is convex and $\mathcal{D} = \mathcal{Y}$. Since $\{f_{1,t}, f_{2,t}, \ldots, f_{N,t}\} \subset \mathcal{D}$ by our assumptions, by convexity $\hat{p}_t \in \mathcal{D}$ as well.

As our goal is to minimize the total regret, it is reasonable to choose the weights according to the regret up to time $t1$. Since $R_{i,t} = L_t - L_{i,t}$, it is the difference between the forecasters total loss and that of expert i after t prediction rounds. The larger $R_{i,t}$ is, the smaller the expert's loss after t rounds is. Therefore, if $R_{i,t}$ is large, we assign a larger weight to expert i and vice versa. Hence, we view the weight as an arbitrary increasing function of the experts regret.

We find it convenient to write this function as the derivative of a nonnegative, convex, and increasing function $\phi : \mathbb{R} \to \mathbb{R}$, whose derivative is denoted as ϕ'. The weighted average forecaster is then defined as

$$\hat{p}_t = \frac{\sum_{i=1}^N \phi'(R_{i,t-1}) f_{i,t}}{\sum_{i=1}^N \phi'(R_{i,t-1})}$$

We start the analysis by making the following observation

Proposition 21.2 *If the loss function $\ell(p, y)$ is convex in the first argument, then*

$$\sup_{y_t \in \mathcal{Y}} \sum_{i=1}^N r_{i,t} \phi'(R_{i,t-1}) \leq 0$$

Proof: By Jensen's inequality we have

$$\ell(\hat{p}, y_t) = \ell\left(\frac{\sum_{i=1}^N \phi'(R_{i,t-1}) f_{i,t}}{\sum_{i=1}^N \phi'(R_{i,t-1})}\right) \leq \frac{\sum_{i=1}^N \phi'(R_{i,t-1}) \ell(f_{i,t}, t)}{\sum_{i=1}^N \phi'(R_{i,t-1})}$$

Rearrange the terms, we obtain the statement above. ■

Based on the proposition above, we can view the weighted average forecaster from another perspective. Define the potential function $\Phi : \mathbb{R}^N \to \mathbb{R}$ of the form

$$\Phi(u) = \psi\left(\sum_{i=1}^N \phi(u_i)\right)$$

where ϕ is any nonnegative, increasing, and twice differentiable function, and ψ is any nonnegative, strictly increasing, concave, and twice differentiable auxiliary function.

Using this notion of potential function, we can give the following equivalent definition of the weighted average forecaster

$$\hat{p}_t = \frac{\sum_{i=1}^N \nabla_i \Phi(R_{i,t-1}) f_{i,t}}{\sum_{i=1}^N \nabla_i \Phi(R_{i,t-1})}$$

. From proposition 21.2 we also know that for any loss function ℓ convex in the first argument,

$$\sup_{y_t \in \mathcal{Y}} \sum_{i=1}^{N} r_{i,t} \nabla_i \Phi(R_{i,t-1}) \leq 0$$

. The above inequality is equivalent to the Blackwell condition. We can then use this condition to prove the following theorem

Theorem 21.3 *Assume that a forecaster satisfies the Blackwell condition for a potential $\Phi : \mathbb{R}^N \to \mathbb{R}$ of the form $\Phi(u) = \psi\left(\sum_{i=1}^N \phi(u_i)\right)$, then for all $n = 1, 2, \ldots$*

$$\Phi(R_n) \leq \Phi(0) + \frac{1}{2}\sum_{t=1}^{n} C_t$$

where

$$C_t = \sup_{u \in \mathbb{R}^N} \left\{ \psi'\left(\sum_{i=1}^{N} \phi(u_i)\right) \sum_{i=1}^{N} \phi''(u_i) r_{i,t}^2 \right\}$$

With Theorem 21.3, we can bound the maximum regret obtained by an arbitrary expert at round n, $\max_i R_{i,n}$, defined as

$$\max_i R_{i,n} = \max_i \sum_{t=1}^{n} (\ell(\hat{p}_t, y_t) - \ell(f_{i,t}, y_t)) = \sum_{t=1}^{n} \left(\ell(\hat{p}_t, y_t) - \min_i \ell(f_{i,t}, y_t)\right)$$

By definitions of ψ and ϕ we have

$$\psi\left(\phi\left(\max_i R_{i,n}\right)\right) = \psi\left(\max_i \phi(R_{i,n})\right) \leq \psi\left(\sum_{i=1}^{N} R_{i,n}\right) = \psi(R_n)$$

. In the definition of the potential function, we also assumed that ψ is invertible. If ϕ happens to be invertible as well, we can easily obtain the following bound

$$\max_i R_{i,n} \leq \phi^{-1}\psi^{-1}(\phi(R_n))$$

.

21.2.1 Example: Polynomially Weighted Average Forecaster

Consider following potential function

$$\Phi(u) = \sum_{i=1}^{N} (u_i)_+^2$$

, where $(u_i)_+ = \max(u_i, 0)$. The weights assigned to the experts are then defined by

$$w_{i,t} = \nabla_i \Phi(R_t) = \partial R_{i,t}$$

, and the regret at round n satisfies

$$\max_i R_{i,n} \leq \sqrt{nN}$$

The proof can be found on page 13 of the recommended reading.

21.2.2 Example: Exponentially Weighted Average Forecaster

Consider following potential function

$$\Phi_\eta(u) = \frac{1}{\eta} \ln \left(\sum_{i=1}^{N} e^{\eta u_i} \right)$$

, where η is a positie parameter. The weights assigned to the experts are of the form

$$\begin{aligned} w_{i,t} = \nabla_i \Phi_\eta(R_t) &= \frac{e^{\eta R_{i,t}}}{\sum_{j=1}^{N} e^{\eta R_{j,t}}} \\ &= \frac{e^{\eta(\sum_{s=1}^{t}(\ell(\hat{p}_s, y_s) - \ell(f_{i,s}, y_s))}}{\sum_{j=1}^{N} e^{\eta(\sum_{s=1}^{t}(\ell(\hat{p}_s, y_s) - \ell(f_{j,s}, y_s))}} \\ &= \frac{e^{-\eta L_{i,t}}}{\sum_{j=1}^{N} e^{-\eta L_{j,t}}} \end{aligned}$$

Using Theorem 21.3, we can show that

$$\max_i R_{i,n} \leq \frac{\ln N}{\eta} + \frac{n\eta}{2}$$

and is minimized when $\eta = \sqrt{\frac{2 \ln N}{n}}$. The proof can be found on page 14 of the recommended reading.