E1 245: Online Prediction & Learning **Fall 2014**

Lecture 6 — August 21

Lecturer: Aditya Gopalan *Scribe: Geethu Joseph*

6.1 Recap

In the last lecture, we looked at the lower bound on the regret or minimax regret, for specific loss function of absolute loss which is not an exp-concave function. We then moved to a specific application called sequential probability estimation or source coding. In this lecture, we show that for sequential probability estimation, with logarithmic loss, we can derive the exact minimax optimum forecaster. We then construct a predictor using ExpWeights algorithm for the special case of *iid* experts and derive the regret bound for the predictor.

6.2 Minimax regret for sequential probability estimation

The minimax regret is defined as, the regret of best prediction algorithm over all possible prediction algorithms and environment outcomes, for a particular choice of loss function and a set of experts. For sequential probability estimation problem, with logarithmic loss, minimax regret is

$$V_T^{(N)} = \inf_{\hat{p}_T} \sup_{y^T \in \mathcal{Y}^T} \log \left[\frac{\sup_{f \in \mathcal{E}} f_T(y^T)}{\hat{p}_T(y^T)} \right],$$

where $\hat{p}_T$ is a possible joint distribution on y^T. The following theorem, due to Shtarkov [1], gives the exact minimax forecaster for the above problem set-up.

Theorem 6.1. *Let $\mathcal{E}$ be a class of experts, then,*

$$\begin{aligned} V_T^{(N)} &= \inf_{\hat{p}_T} \sup_{y^T \in \mathcal{Y}^T} \log \left[\frac{\sup_{f \in \mathcal{E}} f_T(y^T)}{\hat{p}_T(y^T)} \right] \\ &= \sup_{y^T \in \mathcal{Y}^T} \log \left[\frac{\sup_{f \in \mathcal{E}} f_T(y^T)}{p_T^*(y^T)} \right], \end{aligned}$$

where, p_T is the normalized max-likelihood distribution defined as,

$$p_T^*(y_1, y_2, \ldots, y_T) = \frac{\sup_{f \in \mathcal{E}} f_T(y_1, y_2, \ldots, y_T)}{\sum_{\hat{y}^T \in \mathcal{Y}^T} \sup_{f \in \mathcal{E}} f_T(\hat{y}^T)}.$$

Proof: Consider a joint probability distribution p_T, different from the optimal distribution p_T^*. Since $p_T \neq p_T^*$, they should differ for atleast two sequences in $\mathcal{Y}^T$. This is because, we have a constraint that sum of probabilities should add up to unity. Hence, there exist a sequence $y_T \in \mathcal{Y}^T$ such that $p_T^*(y^T) > p_T(y^T)$. Then,

$$\log \left[\frac{\sup\limits_{f \in \mathcal{E}} f_T(y^T)}{p_T(y^T)} \right] > \log \left[\frac{\sup\limits_{f \in \mathcal{E}} f_T(y^T)}{p_T^*(y^T)} \right]. \tag{6.1}$$

Using the definition of p_T^*, we simplify the right hand side of the inequality as

$$\log \left[\frac{\sup\limits_{f \in \mathcal{E}} f_T(y^T)}{p_T^*(y^T)} \right] = \log \left[\sum_{\hat{y}^T \in \mathcal{Y}^T} \sup_{f \in \mathcal{E}} f_T\left(\hat{y}^T\right) \right].$$

Further, we note that the expression has no dependency on the sequence y^T. Thus,

$$\log \left[\frac{\sup\limits_{f \in \mathcal{E}} f_T(y^T)}{p_T^*(y^T)} \right] = \sup_{z^T \in \mathcal{Y}^T} \log \left[\frac{\sup\limits_{f \in \mathcal{E}} f_T(z^T)}{p_T^*(z^T)} \right].$$

Substituting in (6.1), we have, for every $p_T \neq p_T^*$, there exist a sequence $y_T \in \mathcal{Y}^T$, such that

$$\log \left[\frac{\sup\limits_{f \in \mathcal{E}} f_T(y^T)}{p_T(y^T)} \right] > \sup_{z^T \in \mathcal{Y}^T} \log \left[\frac{\sup\limits_{f \in \mathcal{E}} f_T(z^T)}{p_T^*(z^T)} \right].$$

This completes the proof. □

Thus, we have defined the minimax distribution explicitly and the exact conditional probabilities are

$$p_T^*(y|y_1, y_2, \ldots, y_{t-1}) = \frac{p_T^*(y_1, y_2, \ldots, y_{t-1}, y)}{p_T^*(y_1, y_2, \ldots, y_{t-1})}. \tag{6.2}$$

We also point out two major issues associated with the above solution. One, computational complexity in calculating the denominator term in (6.2) increases exponentially with number of rounds. There are N^T terms in the denominator for the T^{th} round. Second, the conditional distribution $p_T^*(.|y_1, y_2, \ldots, y_{t-1}), t = 1, 2, \ldots, T$ highly depends on T, and in general,

$$p_{t-1}^*(y^{t-1}) \neq \sum_{y \in \mathcal{Y}} p^*(y_1, y_2, \ldots, y_{t-1}, y).$$

6.2.1 Construction of predictor: Laplace mixture

In this section, we construct a prediction algorithm for sequential probability estimation problem using ExpWeights algorithm(for the class of all constant experts), assuming *iid* experts. The Laplace mixture forecaster was introduced, in the context of universal coding, by Davisson [2], and also investigated by Rissanen [3]. We note that logarithmic loss function is $\sigma-$exp-concave function, $\forall \sigma \in [0,1]$. We assume that ExpWeights learning rate $\eta = \sigma = 1$. Thus, the the exponentially weighted average forecaster assigns, to each sequence y_t,

$$\begin{aligned}\hat{p}_t(y_t|y^{t-1}) &= \frac{\sum_{f\in\mathcal{E}} f_t(y_t|y^{t-1})e^{-L_f(y^{t-1})}}{\sum_{f\in\mathcal{E}} e^{-L_f(y^{t-1})}} \\ &= \frac{\sum_{f\in\mathcal{E}} f_t(y_t|y^{t-1})f_{t-1}(y^{t-1})}{\sum_{f\in\mathcal{E}} f_{t-1}(y^{t-1})} \\ &= \frac{\sum_{f\in\mathcal{E}} f_t(y^t)}{\sum_{f\in\mathcal{E}} f_{t-1}(y^{t-1})}.\end{aligned}$$

Thus, the joint probability distribution of y^T is given by,

$$\begin{aligned}\hat{p}_T(y^T) &= \prod_{t=1}^{T} \hat{p}_t(y_t|y_{t-1}) \\ &= \frac{\sum_{f\in\mathcal{E}} f_T(y^T)}{\sum_{f\in\mathcal{E}} f_0(y^0)} \\ &= \frac{1}{N}\sum_{f\in\mathcal{E}} f_T(y^T). \end{aligned} \tag{6.3}$$

Thus, it turns out that when there are finite number of experts, probability assigned by the predictor is the the the average of the probabilities assigned by each expert.

This idea may be extended to the case in which the experts are uncountably infinite in numbers. Consider the set of experts as all constant experts, each one of them being all possible *iid* conditional distributions on $\mathcal{Y}$. We model them, using a collection of Bernoulli random variables, with parameter q, where $q \in [0,1]$. Then, we have,

$$f_T(y^T) = q^{n_1}(1-q)^{n_2}, \tag{6.4}$$

and extending (6.3) to this case, we get

$$\hat{p}_T(y^T) = \int_0^1 q^{n_1}(1-q)^{n_2}\mathrm{d}q, \tag{6.5}$$

where n_1 and n_2 are the number of 1's and 0's in the sequence respectively, $n_1 = \sum_{t=1}^{T} y_t$, and $n_2 = T - n_1$. We present a lemma to further simplify the prediction.

Lemma 6.2. *For any two integers, n_1 and n_2,*

$$\int_0^1 q^{n_1}(1-q)^{n_2} dq = \frac{1}{(n_1+n_2+1)\binom{n_1+n_2}{n_1}},$$

where $q \in [0,1]$.

Using the above lemma in (6.5), we have

$$\hat{p}_T(y^T) = \frac{1}{(T+1)\binom{T}{n_1}}. \tag{6.6}$$

Thus, the conditional probability of a t^{th} element of sequence being 1 is given by,

$$\begin{aligned} p(1|y_1, y_2, \ldots, y_{t-1}) &= \frac{p(y_1, y_2, \ldots, y_{t-1}, 1)}{p(y_1, y_2, \ldots, y_{t-1})} \\ &= \frac{\frac{1}{(t+1)\binom{t}{n_1+1}}}{\frac{1}{t\binom{t-1}{n_1}}} \\ &= \frac{n_1+1}{t+1}. \end{aligned} \tag{6.7}$$

Similarly, we get,

$$p(0|y_1, y_2, \ldots, y_{t-1}) = \frac{n_2+1}{t+1}. \tag{6.8}$$

Combining (6.7) and (6.8), we get

$$\hat{p}_t(y|y_1, y_2, \ldots, l_{t-1}) = \frac{1 + \sum_{i=1}^{t-1} \mathbb{I}\{y_i = y\}}{t+1}.$$

The above derived forecaster is called the Laplace mixture forecaster. The rule is also called add-one smoothing, since it includes a one in the numerator so that numerator never vanishes. This prevents the loss function from blowing up to infinity.

Theorem 6.3. *Regret for Laplace mixture is*

$$\sup_{t^T \in 0,1^T} \textit{Regret}_T(\textit{Laplace}) \leq \log(1+T).$$

Proof: Using (6.4) and (6.5) in the definition of minimax regret,

$$\begin{aligned} \text{Regret}_T(\text{Laplace}) &= \sup_{y^T \in \mathcal{Y}^T} \log \left[\frac{\sup_{f \in \mathcal{E}} f_T(y^T)}{\hat{p}_T(y^T)} \right] \\ &= \sup_{y^T \in \mathcal{Y}^T} \sup_{q \in [0,1]} \log \left[\frac{q^{n_1}(1-q)^{n_2}}{\int_0^T q^{n_1}(1-q)^{n_2} \mathrm{d}q} \right]. \end{aligned}$$

Substituting from (6.6),

$$\text{Regret}_T(\text{Laplace}) = \sup_{y^T \in \mathcal{Y}^T} \sup_{q \in [0,1]} \log \left[\frac{q^{n_1}(1-q)^{n_2}}{\frac{1}{(T+1)\binom{T}{n_1}}} \right].$$

To solve the problem of optimization with respect to q, consider the function,

$$\begin{aligned} g(q) &= \log \left[q^{n_1}(1-q)^{n_2} \right] \\ &= n_1 \log(q) + n_2 \log(1-q). \end{aligned}$$

Let $q^* \in [0,1]$ be the unique optimum value that maximizes the concave function $g(q)$. Then, the first derivative of the function vanishes at q^*

$$\frac{\mathrm{d}g}{\mathrm{d}q}(q^*) = 0.$$

This gives,

$$\frac{n_1}{q^*} - \frac{n_2}{1-q^*} = 0,$$

and thus, $q^* = \frac{n_1}{T}$. Substituting back in (6.2.1),

$$\begin{aligned} \text{Regret}_T(\text{Laplace}) &= \sup_{y^T \in \mathcal{Y}^T} \log \left[\left(\frac{n_1}{T}\right)^{n_1} \left(\frac{n_2}{T}\right)^{n_2} \right] + \log \left[(T+1) \binom{T}{n_1} \right] \\ &= \sup_{y^T \in \mathcal{Y}^T} \log \left[\binom{T}{n_1} \left(\frac{n_1}{T}\right)^{n_1} \left(\frac{n_2}{T}\right)^{n_2} \right] + \log(T+1) \\ &\leq \log(T+1). \end{aligned}$$

Last steps follows from the fact that $\log \left[\binom{T}{n_1} \left(\frac{n_1}{T}\right)^{n_1} \left(\frac{n_2}{T}\right)^{n_2} \right] < 0$, since $\binom{T}{n_1} \left(\frac{n_1}{T}\right)^{n_1} \left(\frac{n_2}{T}\right)^{n_2}$ is the probability of getting n_1 heads when a coin is tossed T times, with probability of getting a head in each trail being $\frac{n_1}{T}$. This completes the proof. □

We also add some concluding remarks by stating results for the some other cases:

1. If Beta($\frac{1}{2}, \frac{1}{2}$) is used instead of uniform distribution in $[0,1]$ (which is Beta(1,1) distribution) on *iid* experts, add-1/2 estimator or Krichevsky-Trofimov predictor is obtained. In this case, distribution on q is,

   $$f(q) = \frac{1}{\pi \sqrt{q(1-q)}}.$$

   The prediction algorithm then simplifies to,

   $$\hat{p}_t(1|y^{t-1}) = \frac{\frac{1}{2} + \sum_{s=1}^{t-1} \mathbb{I}\{y_S = 1\}}{t},$$

and regret bound is given by

$$\text{Regret}_T(K-T) \leq \frac{1}{2}\log T + \text{const.}$$

This is the best regret bound that can be obtained for a set of *iid* distribution on experts $\mathcal{Y}$.

2. Furthermore, if the *iid* condition on experts is relaxed to set of all k^{th} order Markov chains on $\mathcal{Y}$, denoted by $\mathcal{E}_k$, the regret bound modifies to

$$\text{Regret}_T(\mathcal{E}_k; \text{add 1/2 rule}) \leq \frac{N^k(N-1)}{2}\log T + \text{const}$$

Bibliography

[1] Y.M. Shtarkov, “Universal sequential coding of single messages”, in *Problems of Information Transmission*, 23:3–17, 1987.

[2] L. D. Davisson, “Universal lossless coding”, in *IEEE Transactions on Information Theory*, 19:783– 795, 1973.

[3] J. Rissanen, “Complexity of strings in the class of Markov sources”, in *IEEE Transactions on Information Theory*, 32:526532, 1986.

[4] Chapter-9, Nicolo Cesa-Bianchi and Gabor Lugosi,“Prediction, Learning and Games”, Cambridge University Press, 2006