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

Lecture 5 — August 19

Lecturer: Aditya Gopalan *Scribe: Raj Kumar Maity*

5.1 Minimax Regret

We showed that for general convex losses

$$R_T(ExpWts) \leq \sqrt{T \log N/2}$$

LOWER BOUND ON REGRET /MINIMAX REGRET:
-Show that without any more structure in the problem(i.e other than convex losses) the bound is UNIMPROVABLE across all algorithm.

5.1.1 Setup

Decision Space D:=[0,1]
Outcome Space y:=$\{0,1\}$
l(p,y):= $|p-y|$
Experts:ξ; $|\xi| = N$
Now consider forecasting algorithm A . We will show that there exists a really bad sequence of outcomes and there exsits a really bad sequence of advice $f_{i,t}$such that

$$\sum_{t=1}^{T} l(\hat{p_t}^A, y_t) - min_{i \in E} \sum_{t=1}^{T} l(f_{i,t}, y_t) \geq \sqrt{T \log N/2}.$$

Proof: Fix $f_{i,t}$

$$\sup_{y_1 \ldots y_T} [\sum_{t=1}^{T} l(\hat{p_t}^A, y_t) - min_{i \in \xi} \sum_{t=1}^{T} l(f_{i,t}, y_t)]$$
$$\geq E_{y_1 \ldots y_T} [\sum_{t=1}^{T} l(\hat{p_t}^A, y_t) - min_{i \in \xi} \sum_{t=1}^{T} l(f_{i,t}, y_t)]$$

[y_i iid Bernoulli r.v (1/2) and $y_i \in \{0,1\}$]

$$
\begin{aligned}
&= \sum_{t=1}^{T} E[l(\hat{p_t}^A, y_t)] - E[min_{i\in\xi} \sum_{t=1}^{T} l(f_{i,t}, y_t)] \\
&= T/2 - E[min_{i\in\xi} \sum_{t=1}^{T} l(f_{i,t}, y_t)] \\
&= E_{Y^T}[max_{i\in\xi} \sum_{t=1}^{T} (1/2 - |f_{i,t} - y_t|)] \\
&= E_{\sigma_1 \ldots \sigma_T}[max_{i\in\xi} \sum_{t=1}^{T} (1/2 - f_{i,t})\sigma_t]
\end{aligned}
$$

(symmetrization)
where$\sigma_t : 1 - 2y_t = \pm 1$ w.p 1/2 (Rademacher random variable)
$[1/2 - |f_{i,t} - y_t| = (1/2 - f_{i,t})(1 - 2y_t)]$
Now

$$
\begin{aligned}
&\sup_{f_{i,t}} \sup_{y_T} REGRET(A) \geq E_{\sigma^T}[max_{i\in\xi} \sum_{t=1}^{T} (1/2 - f_{i,t})\sigma_t] \\
&\geq E_{f_{i,t}\ Ber(1/2)}[E_{\sigma^T}[max_{i\in\xi} \sum_{t=1}^{T} (1/2 - f_{i,t})\sigma_t]] \\
&= 1/2E[max_{i\in\xi} \sum_{t=1}^{T} z_{i,t}\sigma_t] \\
&= 1/2E[max_{i\in\xi} \sum_{t=1}^{T} q_{i,t}] && [z_{i,t} - \textit{iid rademacher r.v] (T,E large)} \\
&= 1/2E[maxG_i]\sqrt{T} && [\text{G}\sim \text{Normal(0,1) iid } \forall i \in \xi] \\
&= 1/2\sqrt{T}\sqrt{2\log|\xi|} \\
&= \sqrt{T\log|\xi|/2}
\end{aligned}
$$

We have used the following well know property of probability theory [1]

1. $\lim_{T\to\infty} E[max_i \frac{\sum_{t=1}^{T} q_{i.t}}{\sqrt{T}}] = E(max_i G_i)$
   [G$\sim$ Normal$(0,1)$ iid $\forall i \in \xi$]
2. $\lim_{T\to\infty} \frac{E[max_i G_i]}{\sqrt{2\log N}} = 1$

SUMMARY:

$$\inf_{alg} \sup_{f_{i,t}} \sup_{y_1..y_t} [REGRET_T(A, y^t, f_{i,t})] \geq \sqrt{T \log N/2}$$

NOTE:

1. Lower bound on $V_T^N \rightarrow$ for any algorithm there exists N experts making prediction and outcomes such that regret(algo)$\geq$ lower bound.
2. Lower bound on $V_T^N \rightarrow$ there exists a good algo for which regret $\leq$upper bound.

□

5.1.2 APPLICATION(Sequential Probability Estimation / Universal Data compression /Universal Source coding)

y={1,2 ... N} (alphabet)
D={N-dim unit simplex} ={all probability distributions on y}
E={Experts} each expert $f \in E$ is a sequence of function f=$(f_1 f_2 \dots)$
$f_t : y^{t-1} \rightarrow D = \Delta_N$
i.e experts' guess distribution of y_t at time t.
$(y_1 \dots y_{t-1}) \mapsto f_t((\cdot)|y_1 y_2 .. y_{t-1}) \in D$

LOSS: log loss $l(p,y) = \sum_{j=1}^N 1(y=j) \log[\frac{1}{p(j)}] = -\log p(j)$
At each time forecaster predicts the distribution
$\hat{p}_t = \hat{p}_t(|y_1 y_{t-1}) \in D$

$$R_T = \sum_{t=1}^T \log \frac{1}{\hat{p}_t(y_t|y_1..y_{t-1})} - \inf_{f \in E} \sum_{t=1}^T \log \frac{1}{f_t(y_t|y_1..y_{t-1})} = \sup_{f \in E} \log[\frac{f_T(y_1..y_T)}{\hat{p_T}(y_1..y_T)}]$$

Regret = Worst case log-likelihood ratio
low regret $\longrightarrow$matches joint probability distribution over sequences

5.1.3 Connection to Information Theory

y^T =MESSAGE$\rightarrow$ENCODER$\rightarrow$ENCODED SEQ.$\rightarrow$DECODER$\rightarrow$MESSAGE
Suppose that y^T was generated by some distribution. encodes y^t into $\leq [-\log_2(f_T(y^T))]$ bits (SHANNON -FANO CODE)
What if:
-the distribution is unknown.
-y_1, y_2.. reveals sequential.
If instead of f_T we used some other distribution g_T. Extra no of bits consumed to encode $(y_1, y_2, .. y_T)$

is

$$\log_2 \frac{1}{g_T(y^T)} - \log_2 \frac{1}{f_T(y^T)} = \frac{1}{\log 2} \log \frac{f_T(y^T)}{g_T(y^T)}$$

Suppose one designed a good sequence of conditional probability estimator $\hat{p}_t(y_t|y_1, y_2..y_{t-1})$ then one can use Arithmatic coding to encode the sequence "on the fly" i.e encode at time t using $\hat{p}_t(y_t|y_1, y_2..y_{t-1})$ such that $\log_2 \hat{p_T}(y_T)$ bits used. So
minimize:$\sup_{f \in \xi} \log \frac{f(y^T)}{p_T(y^T)}$

Bibliography

[1] Nicolo Cesa-Bianchi, Gábor Lugosi, et al. *(Appendix) Prediction, learning, and games*, volume 1. Cambridge University Press Cambridge, 2006.