E1 245: Online Prediction & Learning	Fall 2015

Lecture 3 — August 11

Lecturer: Aditya Gopalan *Scribe: Prakash Barman*

3.1 RECAP

- Math Tools - Probability, Expectation, Variance, Strong Law of Large Numbers, Central Limit Theorem, Markov and Chebyshev's inequality
- Chernoff bound (Hoeffding's inequality). The General case : Let X_1,X_2,...., X_n be iid random variables. Assume X_i are almost surely bounded i.e. $\Pr(X_i \in [a,b]) = 1$, $1 \leq i \leq n$ Then we have $\Pr[\frac{1}{n}\sum_{i=1}^{n} X_i - EX_1 \geq \varepsilon] \leq \exp(\frac{-2n\varepsilon^2}{(b-a)^2})$. If we fix a tolerance ε, then this probability goes down exponentially with n.

In the following sections we continue to see some more tools and concepts that are needed.

3.2 Bernstein's Inequality

Let X_1,X_2,...., X_n be iid random variables with zero mean and $\sigma^2 = Var(X_1)$. Assume $|X_i| \leq 1, \forall i$. Then for all $\varepsilon \geq 0$,

$$\Pr[\frac{1}{n}\sum_{i=1}^{n} X_i > \varepsilon] \leq \exp[\frac{-n\varepsilon^2}{2(\sigma^2 + \frac{\varepsilon}{3})}]. \tag{3.1}$$

The result is useful when variance of X_i is small. Suppose $\sigma^2 \ll 1$ i.e $\sigma^2 = O(\varepsilon) \Rightarrow 2(\sigma^2 + \frac{\varepsilon}{3}) \approx O(\varepsilon)$, then note that the bound on R.H.S $\approx exp(-n\varepsilon)$ vs. $exp(-n\varepsilon^2)$ in Hoeffding's inequality.

3.3 Convexity

3.3.1 Convex Set

A set $K \subseteq \mathbb{R}^d$ is convex, if for any two points that lie in K i.e. $\forall x, y \in K$ and $\forall \lambda \in [0,1]$, the line segment between the two points also lies in K i.e $\lambda x + (1-\lambda)y \in K$

3.3.2 Convex Function

A real-valued function $f : K \to \mathbb{R}$, where $K \subseteq \mathbb{R}^d$ is a convex set, is called convex if the line segment between any two points on the graph of the function lies above the graph. i.e. $\forall x, y \in K$ and $\forall \lambda \in [0,1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \tag{3.2}$$

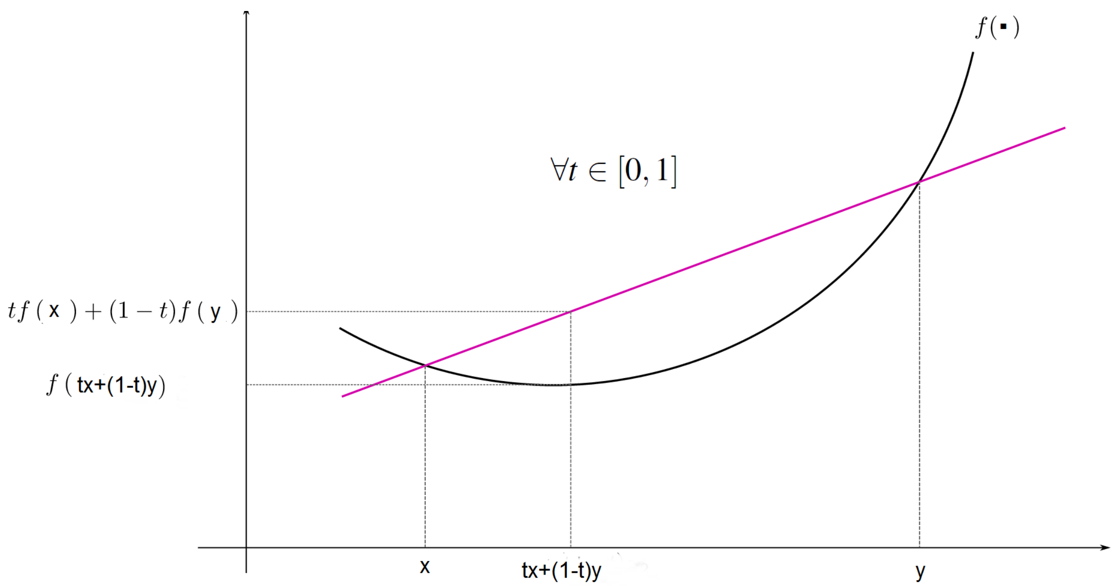

Figure 3.1. convex Function

- $g : K \to \mathbb{R}$ is concave if (-g) is convex

3.3.3 Convex Differentiable Function

Let $K \subseteq \mathbb{R}^d$ be convex. Then the function $f : K \to \mathbb{R}$ is a convex differentiable function if and only if the function lies above all of its tangents i.e. $\forall x, y \in K$,

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) \tag{3.3}$$

The convex differentiable function $f : K \to \mathbb{R}$, $\sigma \geq 0$ is called σ-strongly convex if $\forall x, y \in K$,

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\sigma}{2} \|y-x\|^2 \tag{3.4}$$

If the function f is twice continuously differentiable, then f is strongly convex with parameter σ if and only if $\nabla^2 f(x) \geq \sigma I$ for all x in the domain, where I is the identity and $\nabla^2 f$ is the Hessian matrix. e.g. $f(x) = \frac{\mu}{2}\|x\|^2$ is a μ-strongly convex; $\mu \geq 0$

3.4 Basic Inequalities

3.4.1 Arithmatic-Geometric mean

Given a list of n numbers $\forall x_1, x_2, ..., x_n \geq 0$

$$AM(x_1, x_2, ..., x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \geq GM(x_1, x_2, ..., x_n) = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \tag{3.5}$$

with equality if and only if $\forall i, x_i = x$

3.4.2 Cauchy-Schwarz Inequality

For all vectors $x_1, x_2, ..., x_n; y_1, y_2, ..., y_n \in \mathbb{R}$ it is true that

$$\sum_{i=1}^{n} x_i y_i \leq \sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2} \tag{3.6}$$

with equality if and only if $\exists\ \alpha$ s.t. $y_i = \alpha x_i\ \forall i$

- Notations :
  - Inner Product: $< x, y > = \sum_{i=1}^{n} x_i y_i$
  - Cauchy Schwarz inequality: $< x, y > \leq \|x\|_2 \|y\|_2$

3.4.3 Holder's Inequality

Let $\|x\|_p = \left(\sum_{i=1}^{n} |x|^p\right)^{\frac{1}{p}}$, $p > 0$, $x \in \mathbb{R}^n$ and p & q be atleast 1 i.e. $p, q \geq 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, Then

$$< x, y > \leq \|x\|_p \|y\|_q \tag{3.7}$$

3.4.4 Exponential Inequalities

- $\forall x \in \mathbb{R} : e^x \geq 1 + x$
- $\forall x \geq 0 : e^{-x} \leq 1 - x + \frac{x^2}{2}$
- $\forall x \leq \frac{1}{2} : e^{-x-x^2} \leq 1 - x$

3.5 1 bit prediction with expert advice

Consider a learning protocol where learner makes some predictions at some discrete time instances based on the predictions made by a pool of *experts* and the previous outcomes, e.g.

- Suppose the learner is trying to predict the movement of a stock on the market. Say, on day 't' $y_t \in \{0,1\}$ is the outcome. [e.g. $0 \rightarrow valuedecreased, 1 \rightarrow valueincreased$]
- Additionally you have access to the advice of N "EXPERTS" [Here, EXPERTS can be thought of as financial analysts, or some algorithms/rules, or it may be rumours/news/media etc.]. Let's consider the below prediction model:

Algorithm Prediction model

1: At round t=1,2,3,4,....
2: Observe recommendations of experts: $f_{i,1} \in \{0,1\}\ \forall i \in [N]$
3: Predicted output p_t based on the "recommendations + past information"
4: see actual output y_t

Our Goal is to minimize the number of prediction mistakes $M_T(A)$ made by the algorithm (say Algorithm 'A'). $M_T(A) = \sum_{t=1}^{T} \mathbb{1}\{f_{i,t} \neq y_t\}$, where T is the total number of rounds and A is the prediction algorithm. In the following subsections, we will see some algorithms to achieve our goal. For our next algorithm (Halving/Mjority algorithm), we will assume that there exists a perfect expert which makes no mistakes in predicting the outcomes i.e. $\min_{i \in [N]} M_T(i) = 0$.

3.5.1 Halving/Majority Algorithm [Barzdin and Freivalds (1972), Angluin (1988)]

Algorithm Halving/Majority Algorithm

1: **INPUT:**

Prediction of N experts are available : $f_{1,t}, f_{2,t}, \ldots.., f_{N,t} \in \{0,1\}$

Suppose $\exists$ a perfect expert $j \in [N]$ s.t. $f_{j,t} = y_t\ \forall t \in [T]$, i.e $M_T(j) = 0$

2: **INITIALIZE:**

Trust all experts initially

Algorithm Halving/Majority Algorithm (continued)

3: At round t=1,2,3,4,....

Observe recommendations of experts: $f_{i,1} \in \{0,1\}$ $\forall i \in [N]$

Predicted output p_t = majority (S_t), where $S_1 = [N]$ and $S_t \subseteq [N]$

see actual output y_t

Stop trusting/discard experts that were wrong i.e. $S_{t+1} = \{i \in S_t : f_{i,t} = y_t\}$

4: END

Theorem 3.1. *Under the assumption that there exists a perfect expert, Halving/Majority algorithm will make at most* $log_2 N$ *mistakes,* $M_T(MAJ) \leq log_2 N$

Proof: Observation 1: Whenever Majority makes a mistake - Number of experts reduces by a factor of $\frac{1}{2}$ i.e. $|S_{t+1}| \leq \frac{|S_t|}{2}$. After j^{th} mistake; number of trusted experts $|S_t| \leq \frac{N}{2^j}$.

Observation 2: Because there is a perfect expert, we will always have $|S_T| \geq 1$.

Using observations 1 and 2, we have

$$\frac{N}{2^j} \geq 1$$
$$N \geq 2^j$$
$$log_2 N \geq j \qquad i.e.$$
$$M_T(MAJ) \leq log_2 N$$

□

Homework:

* Show that under the assumption that there exists a perfect expert

$$\max_{y_1 \dots y_T} [M_T(MAJ) - \min_{i \in [N]} M_T(i)] \leq log_2 N$$

this bound is tight. i.e. no other alogrithm performs better;$\forall Algo(A)$ $\exists$ a sequence $y_1 \dots y_T$ and $f_{i,t}$, $i \in [N], t \leq T$ along with perfect expert such that $M_T(A) \geq log_2 N$

* What if number of mistakes by best expert is not zero, i.e. $\min_{i \in [N]} M_T(i) = m \neq 0$

  - Show: A simple modification of majority algorithm gives $M_T(Algo) \leq (m+1) log_2 N$

We will now show an algorithm that gets $M_T(Algo) \leq am + blog_2 N$; for some constants a and b that don't depend on m and N. The idea for designing the algorithm: importance or trust of an expert goes down with number of mistakes.

3.5.2 Weighted Majority Algorithm [Littlestone-Warmuth (1994)]

Algorithm Weighted Majority Algorithm

1: **INPUT:**

Prediction of N experts are available : $f_{1,t}, f_{2,t}, \ldots.., f_{N,t} \in \{0,1\}$

2: **INITIALIZE:**

Initially assign weight 1 to all experts : $w_{i,1} = 1, \forall i \in [N]$

Fix $\varepsilon \in [0,1]$

3: At each round $t \geq 1$

- Observe recommendations of experts: $f_{i,1} \in \{0,1\}\ \forall i \in [N]$
- Predicted output p_t is given by

$$p_t = \begin{cases} 1, & \text{if } \sum\limits_{i:f_{i,t}=1} w_{i,t} \geq \sum\limits_{i:f_{i,t}=0} w_{i,t} \\ 0, & \text{otherwise} \end{cases}$$

- see actual output y_t
- Re-weight each expert: $w_{i,t+1} = w_{i,t}(1-\varepsilon)^q$; $q = \mathbb{1}_{\{f_{i,t} \neq y_t\}}$

4: END

Note :

- For $\varepsilon = 1$: Weighted Majority Algorithm $\equiv$ Halving/MajorityAlgorithm

Theorem 3.2. *For any sequence of instances with binary labels i.e.* $\forall y_1.....y_T \in \{0,1\}^T$ *with N expert predictions avaiable at each round* $f_{i,t}, i \in [N]$ *and* $1 \leq t \leq T$*, with parameter* $\varepsilon \in [0,1]$*, then weighted majority algorithm mistakes at most will be given by*

$$M_T(WTMAJ(\varepsilon)) \leq \frac{(\min\limits_{i \in [N]} M_T(i))log(\frac{1}{1-\varepsilon}) + logN}{log(\frac{1}{1-\frac{\varepsilon}{2}})} \tag{3.8}$$

Proof: Proof will be done in next class □

Corollary 3.3. *if* $\varepsilon \leq \frac{1}{2}$*, then* $M_T(WTMAJ(\varepsilon)) \leq a(m) + b(logN)$*, for some constants a and b that don't depend on m and N; m is number of mistakes by best expert*

Proof: We know that $\forall x \leq \frac{1}{2} : e^{-x-x^2} \leq 1-x \Rightarrow log(\frac{1}{1-\varepsilon}) \leq \varepsilon + \varepsilon^2$
Also, $\forall x \in \mathbb{R} : e^x \geq 1+x \Rightarrow log(\frac{1}{1-\frac{\varepsilon}{2}}) \geq \frac{\varepsilon}{2}$
$\therefore M_T(WTMAJ(\varepsilon)) \leq \frac{m(\varepsilon+\varepsilon^2)+logN}{\frac{\varepsilon}{2}} = 2(1+\varepsilon)m + \frac{2logN}{\varepsilon} = a(m) + b(logN)$ □

References

[1] Nicolo Cesa-Bianchi and Gabor Lugosi, “Prediction, Learning and Games”, Cambridge University Press, 2006.

[2] Gabor Bartok, David Pal, Csaba Szepesvari, and Istvan Szita, “Online learning - CMPUT 654”, Course Notes. 2011.