E1 245: Online Prediction & Learning **Fall 2015**

Lecture 4 — August 13

Lecturer: Dr. Aditya Gopalan *Scribe: Mohammadi Zaki*

4.1 Recap- 1 Bit Prediction Problem

4.1.1 Weighted-Majority Algorithm [Littlestone & Warmuth, '94]

Algorithm : WT-MAJ (ϵ)

---

1. Set parameter $\epsilon \in [0, 1]$ (fixed).
2. Initialize $W_{i,1} = 1, \forall i = [N]$.
3. At each time $t = 1, 2, 3, \ldots$

$$Predict = \begin{cases} 1 & \text{if } \sum_{i;f_{i,t}=1} W_{i,t} \geq \sum_{i;f_{i,t}=0} W_{i,t} \\ 0 & \text{if } otherwise. \end{cases}$$

4. Update $\forall i$,

$$W_{i,t+1} = W_{i,t}(1-\epsilon)^{\mathbf{1}[f_{i,t} \neq y_t]}$$

4.2 Mistake Bound for Weighted-Majority Algorithm (WT-MAJ(ϵ))

Theorem 4.1. ***(Mistake Bound for WT-MAJ)***

$$M_T(WT - MAJ((\epsilon)) \leq \frac{(\min_{i\in[N]} M_T(i))(\log(\frac{1}{1-\epsilon})) + \log N}{\log(\frac{1}{1-\frac{\epsilon}{2}})} \tag{4.1}$$

In particular, if $\epsilon \leq \frac{1}{2}$,

$$\begin{aligned} M_T(WT - MAJ((\epsilon)) &\leq 2(1+\epsilon)(M_T(i^*)) + \frac{2\log N}{\epsilon} \\ &\approx aM_T(\epsilon^*) + b\log N. \end{aligned}$$

Proof: Let's define a POTENTIAL FUNCTION at time t as $\Phi_T = \sum_{i\in[N]} W_{i,t}$, and let i^* be the best expert, i.e., $i^* = argmin_{i\in[N]} M_T(i)$.

At the beginning,

$$\Phi_1 = N. \tag{4.2}$$

At the end,

$$\begin{aligned}\Phi_T &= \sum W_{i,T} \\ &\geq W_{i^*,T} \\ &= W_{i,1}(1-\epsilon)^{M_T(i^*)} \\ &= (1-\epsilon)^{M_T(i^*)}.\end{aligned}$$

If the algorithm makes a wrong prediction at time t,
$\Rightarrow$ At least half the total weight goes down by a factor $(1-\epsilon)$.

$$\Phi_{t+1} \leq \frac{1}{2}\Phi_t + \frac{1}{2}\Phi_t(1-\epsilon) = \Phi_t(1-\frac{\epsilon}{2}). \tag{4.3}$$

$$\begin{aligned}&\leq \Phi_t(1-\frac{\epsilon}{2})^{\mathbf{1}[\hat{y_t}\neq y_t]} \\ &\leq \Phi_1(1-\frac{\epsilon}{2})^{M_T(WT-MAJ)} \\ &= N(1-\frac{\epsilon}{2})^{M_T(WT-MAJ)}.\end{aligned}$$

$$\Rightarrow N(1-\frac{\epsilon}{2})^{M_T(WTMAJ)} \geq (1-\epsilon)^{M_T(i^*)}$$

$$\Rightarrow M_T(WTMAJ) \leq \frac{M_T(i^*)\log(\frac{1}{1-\epsilon}) + \log N}{log(\frac{1}{1-\frac{\epsilon}{2}})}. \tag{4.4}$$

□

4.2.1 Some notes on WT-MAJ

1. WT-MAJ mistake bound implies that $\forall(y_1, y_2, \ldots, y_t)$, and $\forall(f_{i,t})_{i,t}$, and $\epsilon \leq \frac{1}{2}$

$$M_T(WT-MAJ(\epsilon)) - M_T(i^*) \leq (1+2\epsilon)M_T(i^*) + \frac{2\log N}{\epsilon}. \tag{4.5}$$

2. R.H.S. can be "large" if $M_T(i^*)$ is large.

3. This kind of dependence is *unavoidable* due to "mistake penalty being discontinuous" and under any deterministic algorithm.

4.3 General Case - Prediction With Expert's Advice

In general, for any problem of prediction given some expert's advice at each time, we can model the problem given the following,

- Decision Space $\mathcal{D}$.
- Outcome Space $\mathcal{Y}$.
- Loss Function $l : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}^+$.
- Experts : $\mathcal{E}$

Algorithm : General strategy

---

1. At each round $t = 1, 2, 3, \ldots$,
   Environment picks $y_t \in \mathcal{Y}$.
2. Experts give advice $f_{i,t} \in \mathcal{D}, \forall i \in \mathcal{E}$.
3. Decision maker chooses $\hat{p_t} \in \mathcal{D}$ (based on current and past advice and outcomes).
4. Then decision maker sees y_t, suffers loss $l(\hat{p_t}, y_t)$.

4.3.1 Examples

1. 1-*Bit Prediction*

   $\mathcal{D} = \{0, 1\} = \mathcal{Y}$.
   $l(p, y) = \infty[p \neq y]$, ("0-1" loss).
   $\mathcal{E}$ ="Always predict zero", "Always predict one", or any other more complex rules, etc..

2. *"Online Linear Regression"*

   $\mathcal{D} = \{f : \mathbb{R}^d \to \mathbb{R}, f(x) = \langle \mathbf{w}, \mathbf{x} \rangle, \|\mathbf{w}\|_2 \leq 1\}$.
   $\mathcal{E} = \mathcal{D} = \{f : \mathbb{R}^d \to \mathbb{R}, f(x) = \langle \mathbf{w}, \mathbf{x} \rangle, \|\mathbf{w}\|_2 \leq 1\}$.
   $\mathcal{Y} = \mathbb{R}$
   $\mathcal{X} = \mathbb{R}^d$.
   $l(\mathbf{p}, y; \mathbf{x}) = (\langle \mathbf{p}, \mathbf{x} \rangle - y)^2$.

GOAL : Minimize *REGRET* (relative to the best performing expert) regardless of outcomes/advice.

Definition [**REGRET**] : For time horizon T and an expert $i \in \mathcal{E}$, the *REGRET* of a decision making algorithm $\mathcal{A}$ w.r.t. i upto time T is,

$$R_{i,T}(\mathcal{A}) \equiv R_{i,T} = \sup_{\{y_t\}_t, \{f_{i,t}\}_{i\in[N]}, t\in[T]} [\sum_{t=1}^{T} l((\hat{p}_t, y_t)) - \sum_{t=1}^{T} l(f_{i,t}, y_t)].$$

$$Regret : R_T(\mathcal{A}) = \sup_{\{y_t\}_t, \{f_{i,t}\}_{i\in[N]}} [\sum_{t=1}^{T} l((\hat{p}_t, y_t)) - \inf_{i\in\mathcal{E}} \sum_{t=1}^{T} l(f_{i,t}, y_t)].$$

4.3.2 Notes

1. Regret is the *worst case* measure of performance.
   Alternatively, if $\{y_t\}_t$ and $\{f_{i,t}\}_{i\in[N],t\in[T]}$, were stochastic, we could consider AVERAGE-CASE regret, i.e.,

$$\mathbb{E}[(\sum_{t=1}^{T} \mathbf{1}\{\hat{y}_t \neq y_t\} - \inf_{i\in\mathcal{E}} \sum_{t=1}^{T} \mathbf{1}\{f_{i,t} \neq y_t\})].$$

2. *CRITICISM* : WORST-CASE performance measures are too pessimistic ! (i.e., worry about all sequences !).

3. Linearly growing regret with "T" is "BAD" (e.g. 1-bit prediction, $\mathcal{E} = \{$"Always predict zero", "Always predict one", ...$\}$).
   Sub-linear regret is "GOOD".

4.3.3 For the next lecture

Can we obtain sub-linear regret by applying some reasonable constraints on the structure of our prediction problem?