IE613: Online Machine Learning **Jan-Apr 2018**

Lecture 14: Mistake Bound of Winnow Algorithm

Lecturer: M. K. Hanawal *Scribes: Alisha Arora*

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.*

In the last lecture, we discussed about Winnow Algorithm and how its update rule is different from the Perceptron's update rule. In this lecture, we will prove the mistake bound attained by Winnow Algorithm.

14.1 Mistake Bound of Winnow Algorithm

Theorem 14.1 *Let $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), ..., (x_T, y_T)\}$ be an arbitrary sequence. Let $R = \sup_{x_t}\{||x_t||_\infty\}$. Let $\gamma > 0, w^* \in \mathbb{R}^d$ such that $\forall t$, $y_t < w^*, x_t > \ \geq \gamma$, then the mistake bound of Winnow algorithm*

$$M_{Winnow} \leq \frac{||w^*||ln\ d}{\eta\gamma - ||w*||ln\frac{e^{\eta R}+e^{-\eta R}}{2}}\ ,\ \eta > 0$$

Proof: Note that, the mistake bound of Winnow Algorithm is dependent on dimension d.
Let $w_1 = (\frac{1}{d}, \frac{1}{d}, ..., \frac{1}{d}) \in \triangle d$ where,

$$\triangle d = \{x \in \mathbb{R}^d | \sum_{i=1}^{d} x_i = 1, x_i \geq 0\}$$

is called the d-dimensional simplex.

Let us assume that $w^* > 0$.

Let us define Kull-back Leibler(KL) Divergence as:

$$KL(p||q) = \sum_i p_i ln\frac{p_i}{q_i}$$

This measures the similarity between two quantities p and q.
Assume $q = \frac{w^*}{||w^*||} \in \triangle d$ and $w_t \in \triangle d$.

Consider,

$$\begin{aligned} KL(q||w_t) - KL(q||w_{t+1}) &= \sum_{j=1}^{d} q_j ln\frac{q_j}{w_{j,t}} - \sum_{j=1}^{d} q_j ln\frac{q_j}{w_{j,t+1}} \\ &= \sum_{j=1}^{d} q_j ln\frac{w_{j,t+1}}{w_{j,t}} \end{aligned} \tag{14.1}$$

If there is a mistake in round t, then,

$$w_{j,t+1} = \frac{w_{j,t}exp(\eta y_t x_{j,t})}{z_t}$$

Then, eqn 14.1 becomes,

$$
\begin{aligned}
&= \sum_{j=1}^{d} q_j ln \frac{\cancel{w_{j,t}} exp(\eta y_t x_{j,t})}{z_t \cancel{w_{j,t}}} \\
&= \sum_{j=1}^{d} q_j [lnexp(\eta y_t x_{j,t}) - lnz_t] \\
&= \sum_{j=1}^{d} q_j \ (\eta y_t x_{j,t}) - lnz_t (\sum_{j=1}^{d} q_j) \\
&= \eta y_t \sum_{j=1}^{d} q_j \ (x_{j,t}) - lnz_t && (\because \sum_{j=1}^{d} q_j = 1) \\
&= \sum_{j=1}^{d} \frac{w_j^*}{||w^*||} \ (\eta y_t x_{j,t}) - lnz_t && (Substituting \ q_j = \frac{w_j^*}{||w^*||}) \\
&= \frac{\eta}{||w^*||} \underbrace{\sum_{j=1}^{d} w_j^* y_t x_{j,t}}_{} - lnz_t \\
&= \frac{\eta}{||w^*||} y_t < w^*, x_t > - lnz_t \\
&\geq \frac{\eta\gamma}{||w^*||} - lnz_t && (\because y_t < w^*, x_t > \ \geq \gamma)
\end{aligned}
$$

Here, we have established a lower bound on term $KL(q||w_t) - KL(q||w_{t+1})$.
Now,

$$z_t = \sum_{j=1}^{d} w_{j,t} exp(\eta y_t x_{j,t})$$

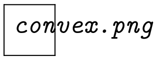
We have assumed, R = $\sup_{x_t} \{||x_t||_\infty\} = \sup_{x_t} \{max(|x_{j,t}| : j = 1, ..., d)\}$.
Hence, each component $x_{j,t}$ is bounded by R. Also,$y_t \in \{-1, 1\}$. Therefore, $y_t x_{j,t} \in [-R, R]$.

Let $\theta = y_t x_{j,t}$. We know, $exp(\eta\theta)$ is a convex function. First, let us define convex functions.

Convex Functions

Definition 14.2 *A function f is said to be convex iff:*

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) \qquad \forall x_1, x_2 \in Dom(f) \text{ and } \forall\ t \in [0,1]$$



This definition does not assume any differentiability condition on f. This definition is called 0^{th} *order characterization* .

Definition 14.3 *A differentiable function f is said to be convex iff:*

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T(x_2 - x_1) \qquad \forall x_1, x_2 \in Dom(f)$$

where, ∇f *is the gradient.*

We know $\theta \in [-R, R]$, therefore θ can be written as a convex combination of -R and R.

$$\theta = t(-R) + (1-t)R \qquad t \in [0,1]$$

Solving this, we get,

$$t = \frac{R-\theta}{2R}$$

We have,

$$\begin{aligned} exp(\eta\theta) &= exp[\eta(t(-R) + (1-t)R)] \\ &\leq texp(-\eta R) + (1-t)exp(\eta R) && (Using\ the\ definition\ of\ convexity) \\ &= \frac{R-\theta}{2R}exp(-\eta R) + \frac{R+\theta}{2R}exp(\eta R) && (Substituting\ t = \frac{R-\theta}{2R}) \\ &= \frac{1-\frac{\theta}{R}}{2}exp(-\eta R) + \frac{1+\frac{\theta}{R}}{2}exp(\eta R) \\ &= [\frac{exp(-\eta R) + exp(\eta R)}{2}] + \frac{\theta}{2R}[\frac{exp(\eta R) - exp(-\eta R)}{2}] \end{aligned}$$

∴ we have,

$$exp(\eta\theta) \leq [\frac{exp(-\eta R) + exp(\eta R)}{2}] + \frac{\theta}{2R}[\frac{exp(\eta R) - exp(-\eta R)}{2}] \tag{14.2}$$

We were considering,

$$
\begin{aligned}
z_t &= \sum_{j=1}^{d} w_{j,t} exp(\eta\theta) \\
&\leq \sum_{j=1}^{d} w_{j,t}([\frac{exp(-\eta R) + exp(\eta R)}{2}] + \frac{\theta}{2R}[\frac{exp(\eta R) - exp(-\eta R)}{2}]) && (using\ 14.2) \\
&= \sum_{j=1}^{d} w_{j,t}[\frac{exp(-\eta R) + exp(\eta R)}{2}] + \frac{1}{R}(\underbrace{\sum_{j=1}^{d} w_{j,t} y_t x_{j,t}})[\frac{exp(\eta R) - exp(-\eta R)}{2}] && (\because \theta = y_t x_{j,t}) \\
&= \sum_{j=1}^{d} w_{j,t}[\frac{exp(-\eta R) + exp(\eta R)}{2}] + \frac{1}{R}(y_t < w^*, x_t >)[\frac{exp(\eta R) - exp(-\eta R)}{2}]
\end{aligned}
$$

We know, when $x \geq 0$,

$$exp(x) - exp(-x) \geq 0$$

Therefore,

$$z_t \leq \sum_{j=1}^{d} w_{j,t}[\frac{exp(-\eta R) + exp(\eta R)}{2}] \qquad (\because\ y_t < w^*, x_t > \ < 0)$$

Now, we considered,

$$KL(q||w_t) - KL(q||w_{t+1})$$

- If there is a mistake, then

$$KL(q||w_t) - KL(q||w_{t+1}) \geq \frac{\eta\gamma}{||w^*||} - ln \sum_{j=1}^{d} w_{j,t}[\frac{exp(-\eta R) + exp(\eta R)}{2}] \tag{14.3}$$

- If there is no mistake,then

$$KL(q||w_t) - KL(q||w_{t+1}) = 0 \qquad (\because w_t = w_{t+1})$$

Summing 14.3 over t=1,...,T , we get

$$KL(q||w_1) - KL(q||w_{t+1}) \geq M(\frac{\eta\gamma}{||w^*||} - ln \sum_{j=1}^{d} w_{j,t}[\frac{exp(-\eta R) + exp(\eta R)}{2}]) \tag{14.4}$$

As,$w_t \in \triangle d$. $\therefore \sum_{j=1}^{d} w_{j,t} = 1$.

Now consider,

$$\begin{aligned} KL[q||w_1] &= \sum_{j=1}^{d} q_j ln \frac{q_j}{w_{j,1}} \\ &= \sum_{j=1}^{d} q_j ln \frac{q_j}{\frac{1}{d}} && (\because w_1 = (\frac{1}{d}, \frac{1}{d}, ..., \frac{1}{d})) \\ &= \sum_{j=1}^{d} q_j ln\ d + \sum_{j=1}^{d} q_j ln\ q_j \\ &= ln\ d \sum_{j=1}^{d} q_j + \sum_{j=1}^{d} q_j ln\ q_j \\ &\le ln\ d && (\because \sum_{j=1}^{d} q_j = 1\ and\ \sum_{j=1}^{d} q_j ln\ q_j \le 0) \end{aligned}$$

$KL[q||w_{T+1}] \geq 0$

$$ln\ d - 0 \geq KL[q||w_1] - KL[q||w_{T+1}] \geq M(\frac{\eta\gamma}{||w^*||} - ln \sum_{j=1}^{d} w_{j,t}[\frac{exp(-\eta R) + exp(\eta R)}{2}]) \qquad (by\ 14.4)$$

$$M \leq \frac{||w^*||ln\ d}{\eta\gamma - ||w*||ln\frac{e^{\eta R}+e^{-\eta R}}{2}}$$

■

Suppose we have knowledge about η & R, we can tighter this bound.
$\therefore$ optimizing over η.(Bound will be min. if denominator is max.).Differentiating denominator w.r.t. η.

$$\gamma - ||w^*|| \frac{Rexp(\eta R) - Rexp(-\eta R)}{\frac{exp(-\eta R)+exp(\eta R)}{2}} = 0$$

we get,

$$\eta^* = \frac{1}{2R} ln \frac{1 + \gamma/(R||w^*||)}{1 - \gamma/(R||w^*||)}$$

Perceptron Update

- If there is a mistake,
$$w_{t+1} = w_t + y_t x_t$$
- If there is no mistake,
$$w_{t+1} = w_t$$

This type of update is subgradient descent using hinge loss which is, $max\{0, -y_t < w, x_t >\}$.
Here, subgradient $\delta l = -y_t x_t$.

Converting $max\{0, -y_t < w, x_t >\}$ to Perceptron update, we get,
$\rightarrow$ *Learner guesses* w_t
$\rightarrow$ *Environment gives* x_t
$\rightarrow w_{t+1} \leftarrow\ w_t - \delta l(x_t)$
This leads us to the idea of General Online Convex Optimization.

14.2 General Online Convex Optimization Problem

In general convex optimization problem, we have:

$$\min_{x \in C} f(x) = f^*$$

where $f : C \to \mathbb{R}$ is convex.

What will happen if we convert this to online setting?

Let us suppose we are given a sequence of convex functions $C_1, C_2, ...$ and points $x_1, x_2, ...$ i.e. in each round, we assume that we are getting convex functions.
Our aim is to find an algorithm A^* which achieves the best possible regret R(A,T):

$$Regret(A,T) = \sum_{t=1}^{T} c_t(x_t) - \min_{x \in C} \sum_{t=1}^{T} c_t(x)$$

where,

$$c_t(x_t) = loss/cost$$

We have to find an algorithm such that:

$$A^* = \arg\min_{A} Regret(A,T)$$