CS 7545: Machine Learning Theory **Fall 2018**

Lecture 9: Zero-sum Game + Boosting

Lecturer: Jacob Abernethy *Scribes: Sheng Zhang*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

9.1 Zero-sum Game

Definition 9.1 (No-regret Algorithm) *An algorithm $\mathcal{A}$ is no-regret if for any sequence $\ell_1, \cdots, \ell_T, \cdots \in [0,1]^n$ with $p_t \in \Delta_n$ chosen as $p_t \leftarrow \mathcal{A}(\ell_1, \cdots, \ell_{t-1})$, satisfies*

$$\epsilon_T \triangleq \frac{1}{T}\left(\sum_{t=1}^{T} p_t^T \ell_t - \min_{p \in \Delta_n} \sum_{t=1}^{T} p^T \ell_t\right) = o(1)$$

Recall that a sequence $a_1, a_2, \cdots$ is $o(1)$ if $\lim_{n\to\infty} a_n = 0$. And note that

$$\min_{p \in \Delta_n} p^T \ell = \min_{i \in \{1,\cdots,n\}} e_i^T \ell$$

where e_i is the standard unit vector with the ith element equal to 1.

Claim: EWA is a no-regret algorithm with

$$\epsilon_T \leq \frac{\log N + \sqrt{2T \log N}}{T} = \frac{\log N}{T} + \sqrt{\frac{2 \log N}{T}}$$

Theorem 9.2 (von Neumann's minimax theorem) *Let $M \in [0,1]^{n\times m}$, then*

$$\min_{p \in \Delta_n} \max_{q \in \Delta_m} p^T M q = \max_{q \in \Delta_m} \min_{p \in \Delta_n} p^T M q$$

Proof: The weak duality

$$\min_{p \in \Delta_n} \max_{q \in \Delta_m} p^T M q \geq \max_{q \in \Delta_m} \min_{p \in \Delta_n} p^T M q$$

is easy to check.
We now prove the strong duality

$$\min_{p \in \Delta_n} \max_{q \in \Delta_m} p^T M q \leq \max_{q \in \Delta_m} \min_{p \in \Delta_n} p^T M q$$

holds.

Let $\mathcal{A}$ be a no-regret algorithm. We will play this game repeatedly!
Protocol:

For $t = 1, 2, \cdots, T$:

- p_t is chosen as $\mathcal{A}(\ell_1, \cdots, \ell_{t-1})$.
- q_t is chosen as $q_t = \operatorname{argmax}_{q \in \Delta_m} p_t^T M q$.

Q1: How happy is the player q after T rounds?

Answer:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} p_t^T M q_t &= \frac{1}{T}\sum_{t=1}^{T} \max_{q\in\Delta_m} p_t^T M q \\
&\geq \max_{q\in\Delta_m} (\frac{1}{T}\sum_{t=1}^{T} p_t)^T M q \\
&= \max_{q\in\Delta_m} \bar{p}^T M q \\
&\geq \min_{p\in\Delta_n} \max_{q\in\Delta_m} p^T M q
\end{aligned} \tag{9.1}$$

Q2: How happy is the player p after T rounds?

Answer:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} p_t^T M q_t &= \frac{1}{T}\sum_{t=1}^{T} p_t^T \ell_t \\
&= \frac{1}{T}\min_{p\in\Delta_n} \sum_{t=1}^{T} p^T \ell_t + \epsilon_T \\
&= \min_{p\in\Delta_n} \frac{1}{T}\sum_{t=1}^{T} p^T M q_t + \epsilon_T \\
&= \min_{p\in\Delta_n} p^T M (\frac{1}{T}\sum_{t=1}^{T} q_t) + \epsilon_T \\
&= \min_{p\in\Delta_n} p^T M \bar{q} + \epsilon_T \\
&\leq \max_{q\in\Delta_m} \min_{p\in\Delta_n} p^T M q + \epsilon_T
\end{aligned} \tag{9.2}$$

It follows from 9.1 and 9.2 that

$$\min_{p\in\Delta_n} \max_{q\in\Delta_m} p^T M q \leq \frac{1}{T}\sum_{t=1}^{T} p_t^T M q_t \leq \max_{q\in\Delta_m} \min_{p\in\Delta_n} p^T M q + 0,$$

(as $\epsilon_T \to 0$ when $T \to \infty$) ■

9.2 Boosting

AdaBoost, short for Adaptive Boosting, is a machine learning meta-algorithm formulated by Yoav Freund and Robert Schapire, who won the 2003 Gdel Prize for their work. Boosting is simply solving a zeor-sum game.

Setup: Suppose we are given n data points $x_1, \cdots, x_n \in \mathcal{X}$, their corresponding labels $y_1, \cdots, y_n \in \{-1, 1\}$ and a set of Hypothesis $\mathcal{H} = \{h_1, \cdots, h_m\}$, where $h_i : \mathcal{X} \to \{-1, 1\}$.

Definition 9.3 (Weak Learning Assumption $(\gamma > 0)$**)** *For any $p \in \Delta_n$, where p_i is the weight for x_i, $\exists\ h \in \mathcal{H}$ satisfying*

$$\mathbb{P}\left[h(x_i) \neq y_i\right] \leq \frac{1}{2} - \frac{\gamma}{2}$$

Note that

$$\begin{aligned}\mathbb{P}[h(x_i) \neq y_i] &= \sum_{i=1}^{n} p_i \cdot \mathbb{1}[h(x_i) \neq y_i] \\ &= \sum_{i=1}^{n} p_i \cdot \left(\frac{1 - h(x_i) y_i}{2}\right)\end{aligned} \tag{9.3}$$

Therefore,

$$\mathbb{P}\left[h(x_i) \neq y_i\right] \leq \frac{1}{2} - \frac{\gamma}{2} \Leftrightarrow \gamma \leq \sum_{i=1}^{n} p_i y_i h(x_i)$$

Definition 9.4 (Strong Learning Hypothesis) *$\exists q \in \Delta_m$, where $m = |\mathcal{H}|$ and q_j is the weight for h_j, such that $\forall i = 1, \cdots, n$,*

$$\left(\sum_{j=1}^{m} q_j h_j(x_i)\right) y_i > 0 \qquad \text{“}q\text{-weighted majority vote of } x_i\text{'s label”}$$

Theorem 9.5 *Boosting via minimax duality.*

Proof: Suppose $\mathcal{H} = \{h_1, \cdots, h_m\}$ satisfies the weak learning assumption, and we are given data $\mathcal{X} = \{x_1, \cdots, x_n\}$. Let $M \in [-1, 1]^{n \times m}$ be a matrix such that

$$M_{ij} = h_j(x_i) y_i$$

Weak Learning Assumption($\gamma > 0$): $\forall p \in \Delta_n$, $\exists j \in [m]$ such that

$$0 < \gamma \leq \sum_{i=1}^{n} p_i y_i h_j(x_i) = p^T M e_j \leq \max_{j \in [m]} p^T M e_j = \max_{q \in \Delta_m} p^T M q \leq \min_{p \in \Delta_n} \max_{q \in \Delta_m} p^T M q$$

By von Neumann's minimax theorem, we have

$$0 < \gamma \leq \min_{p \in \Delta_n} \max_{q \in \Delta_m} p^T M q = \max_{q \in \Delta_m} \min_{p \in \Delta_n} p^T M q = \min_{p \in \Delta_n} p^T M q^* = \min_{i \in [n]} e_i^T M q^*$$

where $q^* \in \underset{q \in \Delta_m}{\operatorname{argmax}} \{\min_{p \in \Delta_n} p^T M q\}$.
Hence, $\forall i = 1, \cdots, n$,

$$\left(\sum_{j=1}^{m} q_j^* h_j(x_i)\right) y_i > 0$$

■