CS 7545: Machine Learning Theory **Fall 2019**

Lecture 10: Boosting and Online Convex Optimization

Lecturer: Jacob Abernethy *Scribes: Arda Pekis, Jinwoo Go*

10.1 Boosting Setting

Assume n examples $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ where $\forall i, y_i \in \{-1, 1\}$ and a set of weak learners $h_1, \cdots, h_m$ where $\forall j, \forall x, h_j(x) \in \{-1, 1\}$.

Definition 10.1 (Weak Learning Condition) *For positive γ:*

$$\forall \vec{p} \in \Delta_n, \exists h_j s.t. \sum_{i=1}^{n} p_i h_j(x_i) y_i \geq 2\gamma$$
$$\iff \Pr_i \left[h_j(x_i) = y_i\right] \geq \frac{1}{2} + \gamma.$$

In other words, for any weighting of the data, there is a weak learner that is correct by at least γ on average.

Definition 10.2 (Strong Learning Condition)

$$\exists \vec{q} \in \Delta_m s.t. \forall i \in [n], F_{\vec{q}}(x_i) = y_i$$
$$\text{where } F_{\vec{q}}(x) = \text{sign}\left(\sum_{j=1}^{m} q_j h_j(x)\right)$$

In other words, there is some weighting of learners such that the weighted majority is correct for all examples.

10.2 Boosting Game

We will represent boosting as a zero-sum game, with payoff matrix M, bounded in $[0, 1]$.

$$\text{Let } M = [y_i h_j(x_i)]_{\substack{i=1,\cdots,n \\ j=1,\cdots,m}}.$$

Lemma 10.3 *The following are equivalent and true under the Weak Learning Condition.*

$$\min_{\vec{p}\in\Delta_n} \max_{\vec{q}\in\Delta_m} \vec{p}^\top M \vec{q} \geq 2\gamma$$
$$\iff \forall \vec{p}, \ \max_{\vec{q}} \vec{p}^\top M \vec{q} \geq 2\gamma$$
$$\iff \forall \vec{p}, \exists \vec{q^*}, \ \vec{p}^\top M \vec{q^*} \geq 2\gamma$$
$$\iff \forall \vec{p} \max_{j=1,\cdots,m} \vec{p}^\top M \vec{e}_j \geq 2\gamma. \qquad \text{(Definition 10.1)}$$

Lemma 10.4 *The following are equivalent and true under the Strong Learning Condition.*

$$\max_{\vec{p}\in\Delta_n} \min_{\vec{q}\in\Delta_m} \vec{p}^\top M \vec{q} > 0$$
$$\iff \exists \vec{q^*}, \ \min_{\vec{p}} \vec{p}^\top M \vec{q^*} > 0$$
$$\iff \exists \vec{q^*}, \forall \vec{p}, \ \vec{p}^\top M \vec{q^*} > 0$$
$$\iff \exists \vec{q^*}, \forall i \in [n], \ \vec{e}_j^\top M \vec{q^*} > 0. \qquad \text{(Definition 10.2)}$$

Minimax Generally, $\exists \vec{p}$ and $\forall \vec{q}$ commute when minimizing $\vec{p}$ and maximizing $\vec{q}$ and the expression is convex in $\vec{p}$ and concave in $\vec{q}$. In our case, the expression is bi-linear.

Theorem 10.5 *The Weak Learning Condition is equivalent to the Strong Learning Condition.*

Proof: Since the Weak Learning Condition is equivalent to

$$\min_{\vec{p}\in\Delta_n} \max_{\vec{q}\in\Delta_m} \vec{p}^\top M\vec{q} \geq 2\gamma, \qquad \text{(Lemma 10.3)}$$

and the Strong Learning Condition is equivalent to

$$\max_{\vec{p}\in\Delta_n} \min_{\vec{q}\in\Delta_m} \vec{p}^\top M\vec{q} > 0, \qquad \text{(Lemma 10.4)}$$

by the minimax theorem

$$\min_{\vec{p}\in\Delta_n} \max_{\vec{q}\in\Delta_m} \vec{p}^\top M\vec{q} = \max_{\vec{p}\in\Delta_n} \min_{\vec{q}\in\Delta_m} \vec{p}^\top M\vec{q} \geq 2\gamma > 0.$$

Therefore, the Weak Learning Condition is equivalent to the Strong Learning Condition. ■

10.3 Solving Boosting

To solve the boosting game, we need to find a minimax point. Then, to ϵ-approximate the minimax point, $(\vec{\hat{p}}, \vec{\hat{q}}) \in \mathbb{R}^n \times \mathbb{R}^m$ need to satisfy:

$$\max_{\vec{q}} \vec{\hat{p}}^\top Mq \leq V^* + \epsilon$$
$$\min_{\vec{p}} \vec{p}^\top M\vec{\hat{q}} \leq V^* - \epsilon$$

where $V^* = \min_{\vec{p}} \max_{\vec{q}} \vec{p}^\top M\vec{q}$.

For the boosting game, if we find a $\vec{\hat{p}}$, $\vec{\hat{q}}$ which are ϵ- approximation Nash equilibrium and $\epsilon < 2\gamma$ then, for all i : $F_{\hat{q}}(x_i) = y_i$ ($\because$ by assumption)

$$\begin{aligned} \forall i,\ e_i^\top M\vec{q} &\geq V^* - \epsilon \\ &\geq 2\gamma - \epsilon \text{ (WLC)} \\ &\geq 0 \text{ (By assumption)} \qquad (10.1) \\ e_i^\top M\vec{q} &= y_i \Sigma_j \vec{q_j} h_j(x_i) > 0 \\ \iff F_q(x_i) y_i & \end{aligned}$$

$*$ $F_q(x) = \text{sign}(\Sigma_j(\vec{q_j} h_j(x)))$, $F_q(x) = y \iff y = \text{sign}(\Sigma_j \vec{q_j} h_j(x)) \iff y\Sigma h_j(x)\vec{q_j} > 0$.

Algorithm 1 Solving 0-sum game using EWA

for t = 1, ... , T **do**
  $p_t(i) = \exp(-\eta \Sigma_{s=1}^{t-1} e_i^\top M\vec{q_s})/Z$ (i = 1, ..., n)
  $q_t(j) = \exp(-\eta \Sigma_{s=1}^{t-1} \vec{p_s}^\top M e_j)/Z'$ (j = 1, ..., m)
end for
Return $\vec{\hat{p}}, \vec{\hat{q}} = (\frac{1}{T}\Sigma_{t=1}\vec{p_t}, \frac{1}{T}\Sigma_{t=1}\vec{q_t})$

Theorem 10.6 $\vec{\hat{p}}, \vec{\hat{q}}$ *are an* ϵ*-approximation Nash Equilibrium, where* $\epsilon = \frac{Reg_T^p + Reg_T^q}{T}$.

Algorithm 2 In Boosting Game

for t = 1, ... , T **do**
  $p_t(i) = \exp\left(-\eta \Sigma_{s=1}^{t-1} y_i h_{js}(x_i)\right)/Z$ (i = 1, ..., n)
  $q_t(j) = \arg\max_{e_j} \vec{p}^\top M e_j (= \Sigma p(i) h_j(x_i)) = e_{\text{best weak learner at t}}$
end for
Return $\hat{\vec{q}} = (\frac{1}{T}\Sigma_{t=1}\vec{q_t} = \frac{1}{T}\Sigma_{t=1}e_{jt})$

Corollary 10.7 *Consider new version 2.0, when the second person knows $\vec{p}$. $\vec{q_t} = \arg\max_{\vec{q}} \vec{p_t}^\top M\vec{q}, Reg_T^q \leq 0$ (Because the q player plays the best response at every step, the algorithm cumulative loss till time T is lower than the loss of the best fixed action in hindsight). Now $\epsilon = \frac{Reg_T^p}{T}$.*

Q. How many round of Boosting need to ensure perfect training accuracy?

$$\begin{aligned}(\because \text{WLC})\ 2\gamma > \epsilon = \frac{\text{Reg}_T^p}{T} &\leq \frac{\log(N) + \sqrt{T\log(N)}}{T} \\ \iff T > \frac{\log(N)}{4\gamma^2}&.\end{aligned} \tag{10.2}$$