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

Lecture 10: Boosting and Online Convex Optimization

Lecturer: Jacob Abernethy *Scribes: Hongyu Ouyang and Kyle Kosic*

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

10.1 Boosting

Boosting Setup

- Input space $\mathcal{X}$, labels $\mathcal{Y} \in \{-1, 1\}$
- Given "weak" hypothesis $\mathcal{H} : |\mathcal{H}| = m$ where $h \in \mathcal{H}$ is a map $\mathcal{X} \to \mathcal{Y}$
- Given data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, \quad i = 1, ..., n$

Weak Learning Assumption (γ)

$$\forall p \in \Delta_n \quad \exists h \in \mathcal{H} : \sum_{i=1}^{n} p(i) h_p(x_i) y_i \geq \gamma \quad \text{where } \gamma > 0$$

$$\text{equivalent to: } \Pr_{i \sim p}(h(x_i) = y_i) \geq \frac{1}{2} + \frac{\gamma}{2}$$

So h_i is slightly better than random by $\frac{\gamma}{2}$ where γ is usually very small.

Strong Learning Assumption

$$\exists q \in \Delta_m \forall i \in 1, ..., n : \sum_{j=1}^{m} q(j) h_j(x_i) y_i > 0$$

So the majority-weighted vote on each (x_i, y_i) is always correct.

Boosting as a Game Define a game matrix $M \in \{0, 1\}^{n \times m}$ where $M_{ij} = h_j(x_i) y_i$. Then

$$\begin{aligned} \text{WLA}(\gamma) &\iff \min_{p \in \Delta_n} \max_j p \cdot M e_j \geq \gamma \\ &\iff \min_{p \in \Delta_n} \max_{q \in \Delta_m} p \cdot M q \geq \gamma \end{aligned}$$

Additionally

$$\begin{aligned} \text{SLA} &\iff \max_{q \in \Delta_m} \min_{i=1,...,n} e_i \cdot M q > 0 \\ &\iff \max_{q \in \Delta_m} \min_{p \in \Delta_n} p \cdot M q > 0 \end{aligned}$$

By Von Neumann minimax theorem, WLA(γ) $\implies$ SLA, we just need to find q. Consider the iterative process:

Algorithm 1 EWA no-regret

---

for $t = 1...T$ **do**
  p_t chosen by EWA(η)
  $q_t = \arg\max_q p_t \cdot Mq$ (q_t is the best response)
  p player receieves loss vector $\ell_t := Mq_t$
end for
Let $(\bar{p}_T, \bar{q}_T) = \frac{1}{T}\sum_{t=1}^{T}(p_t, q_t)$

Note: this is almost the optimal solution, will be somewhere between v^* (value of the game) and $\bar{R}_T$. Let $\bar{R}_T$ be any regret upper bound on

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

Which we previously showed for EWA was

$$\bar{R}_T = \sqrt{\frac{2 \log N}{T}} + \frac{\log N}{T} \quad \text{where the } \frac{\log N}{T} \text{ is negligible}$$

We also showed

$$\begin{aligned}
v^* \text{ (value of the game)} &= \min_p \max_q p \cdot Mq \\
&\leq \max_q \frac{1}{T} \sum_{t=1}^{T} p_t \cdot Mq \\
&\leq \frac{1}{T} \sum_{t=1}^{T} \max_q p_t \cdot Mq \\
&= \frac{1}{T} \sum_{t=1}^{T} p_t \cdot Mq_t (\text{ best response } q_t) \\
&= \frac{1}{T} \sum_{t=1}^{T} p_t \cdot \ell_t \\
&\leq \min_p \frac{1}{T} \sum_{t=1}^{T} p \cdot \ell_t + \bar{R}_t && \text{(regret bound)} \\
&= \min_p p \cdot M\bar{q}_T + \bar{R}_t \\
&\leq \max_q \min_p p \cdot Mq + \bar{R}_T
\end{aligned}$$

Consider the value of the boosting game:

$$\begin{aligned}
\gamma \leq v^* &\leq \min_p p \cdot M\bar{q}_T + \bar{R}_T \\
&= \min_{i=1,\dots,n} e_i \cdot M\bar{q}_T + \bar{R}_T
\end{aligned}$$

$$\implies \forall i \in 1...n : \sum_{j=1}^{m} \bar{q}_T(j) h_j(x_i) y_i \geq \gamma - \bar{R}_T$$

We need $\gamma - \bar{R}_T > 0$:

$$\bar{R}_T \leq \gamma \iff \sqrt{\frac{2\log N}{T}} < \gamma$$
$$\iff T > \frac{2\log N}{\gamma^2}$$

If you run this algorithm at least $\frac{2\log N}{\gamma^2}$ rounds, then your classification will be correct.

Boosting by Majority Brief summary of the Boosting by Majority algorithm:

Algorithm 2 Boosting by Majority

Let $T > \frac{2\log n}{\gamma^2}$
Let $w_1 + ... + w_N = 1$
for $t = 1...T$ **do**
  $p_t := \frac{w_t}{\sum_{i=1}^N w_t(i)}$
  $h_t = \arg\max_{h \in \mathcal{H}} \sum_{i=1}^N p_t(i)h(x_i)y_i$
  $w_{t+1}(i) = w_t(i)\exp(-\eta h_t(x_i)y_i)$
end for
Output $\bar{h}_T = \frac{1}{T}\sum_{t=1}^T h_t$

In short, we are taking data point i, weighting it more if classified wrong, and weighting it less if classified correctly.

10.2 Online Convex Optimization

Generalized Experts Setting Then consider the following process:

Algorithm 3 Generalized Experts

Let $K \subseteq \mathbb{R}^d$ convex and compact.
for $t = 1...T$ **do**
  Algorithm selects $x_t \in K$
  Nature selects loss convex function $f_t : K \to \mathbb{R}$
end for
$\text{Regret}_T := \sum_{t=1}^T f_t(x_t) - \min_{x\in K}\sum_{t=1}^T f_t(x)$

To see that this is the generalization of the experts setting, let $K = \Delta_n$ and $f_t(x) = \ell \cdot x$.

Online Gradient Descent Use the Generalized Experts setting to define a new algorithm. Let:

$$x_0 = \text{arbitrary point in } K$$

$$x_{t+1} = \text{proj}_K(x_t - \eta\nabla f_t(x_t))$$

Where $\text{proj}_K(x) = \arg\min_{y\in K} \|y - x\|_2$. Note that for any convex set K, for any $z \in K$, and any y, we have the following:

$$\|\text{proj}_K(y) - z\|_2 \leq \|y - z\|_2$$

The closest point in K to y is $\text{proj}_K(y)$ when K is convex.

Theorem 10.1 *Let $\nabla_t = \nabla f_t(x_t)$. Assume $\|\nabla_t\|_2 \leq G$ where G is some constant, and $\|x_0 - x^*\|_2 \leq D$ for any $x^* \in K$ where D is some constant. Then:*

$$R_T(GD) \leq GD\sqrt{T}$$

Proof:

$$\begin{aligned}
\frac{1}{2}\|x_{t+1} - x^*\|_2^2 &= \frac{1}{2}\|\mathrm{proj}_K(x_t - \eta\nabla_t) - x^*\|_2^2 \\
&\leq \frac{1}{2}\|(x_t - x^*) - \eta\nabla_t\|_2^2 \\
&= \frac{1}{2}\|x_t - x^*\|_2^2 + \frac{\eta^2}{2}\|\nabla_t\|_2^2 - \eta\nabla_t \cdot (x_t - x^*) \\
\implies \nabla_t \cdot (x_t - x^*) &\leq \frac{1}{2\eta}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\eta}{2}\|\nabla_t\|_2^2
\end{aligned}$$

Then

$$\begin{aligned}
\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) &\leq \sum_{t=1}^{T} \nabla_t \cdot (x_t - x^*) && \text{(convexity of f)} \\
&\leq \sum_{t=1}^{T} \frac{1}{2\eta}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} + x^*\|_2^2\right) + \underbrace{\frac{\eta}{2}\|\nabla_t\|_2^2}_{\leq G^2} \\
&\leq \frac{1}{2\eta}\left(\underbrace{\|x_1 - x^*\|_2^2}_{\leq D^2} - \underbrace{\|x_{T+1} - x^*\|_2^2}_{0}\right) + \frac{T\eta G^2}{2} \\
&\leq \frac{D^2}{2\eta} + \frac{T\eta G^2}{2} \\
&= GD\sqrt{T} && \text{when } \eta = \frac{D}{G\sqrt{T}}
\end{aligned}$$

■