IE613: Online Machine Learning **Jan-Apr 2016**

Lecture 15: Online Gradient Descent

Lecturer: Palaniappan Balamurugan *Scribes: Aakash Banik*

15.1 Recapitulation

Earlier we were introduced to the perceptron algorithm which is given as:

Algorithm 1: Perceptron algorithm

Input : Linear Hypothesis Set $\{w : w \in \mathbb{R}^d\}$
Output: Predictions for the labels
initialize w=w_0;
for *t in range* $1, \ldots, T$ **do**
  Player receive x_t;
  predict p_t=sign($< w_t, x_t >$)
  environment reveals the loss: $l_t(y_t, p_t)$
  if *a mistake is made* **then**
    Perceptron Update happens
  end
  else
    No Update
  end
end

We want this sort of an algorithm on a general setting.

15.2 General Setting

The Online Gradient Descent Algorithm is given as:

Algorithm 2: Online Gradient Descent

Input : Linear Hypothesis Set $\{w : w \in \mathbb{R}^d\}$
Output: Predictions for the labels
initialize $w_1 \in K$ (The structure of K is specified);
for *t in range* $1, \ldots, T$ **do**
  Player receive x_t;
  predict p_t=sign($< w_t, x_t >$)
  environment reveals a convex loss function: $c_t(y_t, p_t)$
  Player Updates: $w_{t+1} \leftarrow Proj_k(w_t - y_t \nabla c_t)$
end

It is to be noted that in our Online Gradient Descent algorithm, the update happens in all rounds. We assume that K is compact (i.e. closed and bounded)and c_t is differentiable. Also, the gradient of the convex function in the updation step is computed with respect to w_t.

Now, In a convex optimization setting, we have an objective function which we want to optimise over a given set.

$$\underset{u \in K}{minimize} \quad F(u)$$

Here, K is a convex set, i.e. $if \ ,x \in K \ \ and \ \ y \in K, \Rightarrow \lambda x + (1-\lambda)y \in K \ \ \forall \lambda \in [0,1]$.

In a general convex optimization setting we do the general **Gradient Descent**, in which the updation is given as:

Start with u_0
For t=0,1,2,... ,do
$\quad \tilde{u}_{t+1} \leftarrow u_t - y_t \nabla F(u_t)$
$\quad u_{t+1} \leftarrow Proj_K(\tilde{u}_{t+1})$

where the projection function is defined as: $Proj_K(z) = \underset{u \in K}{argmin} \quad ||u - z||_2$

15.3 Regret of Online Gradient Descent Algorithm

Now, we look into the Regret Function of our Online Gradient Descent Algorithm.

$$Regret_{OGD}(u,T) \quad = \quad \sum_{t=1}^{T} c_t(w_t) \quad - \sum_{t=1}^{T} c_t(u)$$

where, $u \in K$
The Realizability assumption vacuously holds in this case, since F(.) is continuous and K is compact, which implies that minima exists [Weierstrass Theorem].
i.e. $\underset{u \in K}{min} \quad \sum_{t=1}^{T} c_t(u) =: u^*$ exists.

Proof:

We had assumed that c_t is convex and differentiable. So,
$c_t(u^*) \geq c_t(w_t) - < \nabla c_t(w_t), w_t - u^* >$
$\Rightarrow c_t(w_t) - c_t(u^*) \ \leq \ < \nabla c_t(w_t), w_t - u^* >$

We claim : $||w_{t+1} - u^*||^2 \ \leq \ ||\tilde{w}_{t+1} - u^*||^2$
The claim can be backed since $w_{t+1} = Proj_K(\tilde{w}_{t+1})$
which implies, $||z - w_{t+1}||^2 \leq ||z - \tilde{w}_{t+1}||^2 \quad \forall \ z \in K$

Then,

$$\begin{aligned}||w_{t+1} - u*||^2 \;&\leq\; ||\tilde{w}_{t+1} - u^*||^2 \\ &= ||w_t - \eta_t \nabla c_t(w_t) - u^*||^2 \\ &= ||w_t - u^*||^2 + \eta_t^2 ||\nabla c_t(w_t)||^2 - 2\eta_t < \nabla c_t(w_t), w_t - u^* >\end{aligned}$$

Lets also assume, $||\nabla c_t(w_t)|| \leq G \quad \forall\, t$
and, $distance(x, y) \leq D \quad \forall\;\; x, y \in K$

We can say,

$$\begin{aligned}2\eta_t < \nabla c_t(w_t), w_t - u^* > \;&\leq\; ||w_t - u^*||^2 - ||w_{t+1} - u*||^2 + \eta_t^2 ||\nabla c_t(w_t)||^2 \\ \Rightarrow < \nabla c_t(w_t), w_t - u^* > \;&\leq\; \frac{1}{2\eta_t}\big[||w_t - u^*||^2 - ||w_{t+1} - u^*||^2 + \eta_t^2 ||\nabla c_t(w_t)||^2\big] \\ \Rightarrow c_t(w_t) - c_t(u^*) \;&\leq \frac{1}{2\eta_t}\big[||w_t - u^*||^2 - ||w_{t+1} - u^*||^2\big] + \frac{\eta_t}{2} G^2\end{aligned}$$

Summing over t=1,2,...,T:

$$\begin{aligned}\sum_{t=1}^{T}\big[c_t(w_t) - c_t(u^*)\big] &\leq \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\eta_t}\big[||w_t - u^*||^2 - ||w_{t+1} - u^*||^2\big] \\ &= \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \frac{1}{2\eta_1}||w_1 - u^*||^2 - \frac{1}{2\eta_{T+1}}||w_{T+1} - u^*||^2 + \frac{1}{2}\sum_{t=2}^{T}\big(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\big)||w_t - u^*||^2 \\ &\leq \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \frac{1}{2\eta_1}||w_1 - u^*||^2 + \frac{1}{2}\sum_{t=2}^{T}\big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\big)||w_t - u^*||^2 \\ &\leq \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \frac{1}{2\eta_1}D^2 + \frac{1}{2}\sum_{t=2}^{T}\big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\big)D^2 \\ &= \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + D^2\big[\frac{1}{2\eta_1} - \frac{1}{2\eta_1} + \cdots + \frac{1}{2\eta_T}\big] \\ &= \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + D^2\frac{1}{2\eta_T}\end{aligned}$$

Now, we would be imposing the assumption that $\eta_t = \frac{1}{t}$

$$\begin{aligned}\Rightarrow Regret_{OGD}(u^*, T) &\leq \frac{G^2}{2}\sum_{t=1}^{T}\frac{1}{t} + \frac{D^2}{2}T \\ &\leq \frac{G^2}{2}\big[1 + \log T\big] + \frac{D^2}{2}T\end{aligned}$$

The above step follows from the approximation that: $\sum_{t=1}^{T}\frac{1}{t} \;\leq\; 1 + \int_0^T \frac{1}{t}dt \;\leq 1 + \log T$

Now, in order to get a better bound, we make the assumption of strong convexity. Strong convexity of a function implies that its curvature is strong and the Hessian matrix $H > 0$. The idea behind assuming a function to be strongly convex is that we want its growth to be quadratic.

First Order characterization of a strongly convex function:

$$f(y) \;\geq\; f(x) + < \nabla f(x), y - x > + \frac{\alpha}{2}||y - x||^2$$

where, α (> 0) is the modulus of strong convexity.

If,f is twice differentiable, then:

$$\nabla^2 f \geq \alpha I$$

If a function $f(x)$ is strongly convex, then the function defined by the quantity $g(x) = f(x) - \frac{\alpha}{2}||x||^2$ is also convex.

So, we assumed that c_t is strongly convex with modulus α. Also, earlier we assumed Realizability with u^*, but we did not assume anything about the uniqueness of u^*. Here, with strong convexity, the minimizer is unique.

Now again, $c_t(u^*) - c_t(w_t) \;\geq\; < \nabla c_t(w_t), u^* - w_t > + \frac{\alpha}{2}||u^* - w_t||^2$

Again, we claim:

$$\begin{aligned}
||w_{t+1} - u^*||^2 &\leq ||\tilde{w}_{t+1} - u^*||^2 \\
&= ||w_t - u^*||^2 + \eta_t^2||\nabla c_t(w_t)||^2 - 2\eta_t < \nabla c_t(w_t), w_t - u^* >
\end{aligned}$$

Therefore, we get:

$$c_t(w_t) - c_t(u^*) \;\leq\; \frac{1}{2\eta_t}\big[||w_t - u^*||^2 - ||w_{t+1} - u^*||^2\big] + \frac{\eta_t}{2}G^2 - \frac{\alpha}{2}||u^* - w_t||^2$$

Summing over t=1,2,...,T:

$$\begin{aligned}
\sum_{t=1}^{T}\big[c_t(w_t) - c_t(u^*)\big] &\leq \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \sum_{t=1}^{T}\big(\frac{1}{2\eta_t} - \frac{\alpha}{2}\big)\big[||w_t - u^*||^2 - \frac{1}{2\eta_t}||w_{t+1} - u^*||^2\big] \\
&= \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \sum_{t=1}^{T}\big[\frac{1}{2\eta_t}||w_t - u^*||^2 - \frac{1}{2\eta_t}||w_{t+1} - u^*||^2 - \frac{\alpha}{2}||w_t - u^*||^2\big] \\
&= \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \frac{1}{2\eta_1}||w_1 - u^*||^2 - \frac{1}{2\eta_T}||w_{T+1} - u^*||^2 - \frac{\alpha}{2}||w_1 - u^*||^2 \\
&\quad + \frac{1}{2}\sum_{t=2}^{T}\big[\big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\big)||w_t - u^*||^2 - \alpha||w_t - u^*||^2\big] \\
&\leq \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \big(\frac{1}{2\eta_1} - \frac{\alpha}{2}\big)D^2 + \frac{1}{2}\sum_{t=2}^{T}\big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \alpha\big)D^2
\end{aligned}$$

Now, if we assume $\eta_t = \frac{1}{\alpha_t}$, then:

$$\sum_{t=1}^{T}\big[c_t(w_t) - c_t(u^*)\big] = \frac{G^2}{2\alpha^2}\sum_{t=1}^{T}\frac{1}{t} + \big(\frac{\alpha}{2} - \frac{\alpha}{2}\big)D^2 + \frac{1}{2}\sum_{t=2}^{T}\big(\alpha t - \alpha(t-1) - \alpha\big)D^2$$
$$\leq \frac{G^2}{2\alpha}\big(\log T + 1\big)$$

which is a better regret as compared to the previous case.

15.4 Lower Bound for Perceptron-type Algorithms

Theorem 15.1 *Let,* $\mathbb{X} = \{x \in \mathbb{R}^d \ \mid ||x|| \leq 1\}$. *Also let,* $\frac{1}{\gamma^2} \leq d$.
Then for any deterministic algorithm A*, there exists a dataset which is separable by margin* γ *on which* A *makes atleast* $\left[\frac{1}{\gamma^2}\right]$ *mistakes.*

Proof
Construction of the Dataset

Assume, $n = \left[\frac{1}{\gamma^2}\right]$
$\quad \Rightarrow n \leq d \quad since, \ \frac{1}{\gamma^2} \leq d$
Also, $\gamma^2 n \leq 1$
Let e_i be the i^{th} standard basis vector. Also, let b denote the vector of labels for n data points.
Then, $\forall \ \ b \in \{-1,1\}^n, \ \ \exists$ a $w \in X \ni ||w|| \leq 1 \quad and \ \ b_i(w_i e_i(i)) = \gamma \quad \forall \ \ i = 1(1)n$
$\Rightarrow b_i w_i 1 = \gamma$
$\Rightarrow w_i = \frac{\gamma}{b_i}$
$\Rightarrow w_i = \gamma b_i \quad since, \ b_i \in \{-1,1\}$

So, we can come up with such a w by taking $w_i = \gamma b_i$, when i=1(1)n; and 0, otherwise.
Now,

$$\begin{aligned} ||w||^2 &= w_1^2 + \cdots + w_d^2 \\ &= w_1^2 + \cdots + w_n^2 \quad since, w_{n+1} = \cdots = w_d = 0 \\ &= \gamma^2 b_1^2 + \cdots + \gamma^2 b_n^2 \\ &= \gamma^2\big[b_1^2 + \cdots + b_n^2\big] \\ &= \gamma^2 n \\ &\leq 1 \end{aligned}$$

Therefore, $w \in X$.

For any algorithm A, let $x_i = e_i \ \ \forall \ \ i = 1(1)n$
Set, $y_1 = -A(x_1)$
$\quad y_2 = -\big[A((x_1, y_1), x_2)\big]$
$\quad y_3 = -\big[A((x_1, y_1), (x_2, y_2), x_3\big]$
$\quad \vdots$
Remark: For these n data points, we force the algorithm to make mistakes on all sample points. We see that n mistakes are made even when the data is separable by a margin γ.