Lecturer: Rob Schapire
Scribe: Shaoqing Yang

Lecture # 18
April 10, 2014

1 Widrow-Hoff Algorithm

First let's review the Widrow-Hoff algorithm that was covered from last lecture:

Algorithm 1: Widrow-Hoff Algorithm

Initialize parameter $\eta > 0$, $\mathbf{w}_1 = \mathbf{0}$
for $t = 1 \ldots T$
  get $\mathbf{x}_t \in \mathbb{R}^n$
  predict $\hat{y}_t = \mathbf{w}_t \cdot \mathbf{x}_t \in \mathbb{R}$
  observe $y_t \in \mathbb{R}$
  update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta(\mathbf{w}_t \cdot \mathbf{x}_t - y_t) \cdot \mathbf{x}_t$

And we define the loss functions as $L_A = \sum_{t=1}^{T}(\hat{y}_t - y_t)^2$. And $L_{\mathbf{u}} = \sum_{t=1}^{T}(\mathbf{u} \cdot \mathbf{x}_t - y_t)$. What we want is

$$L_A \leq \min_{\mathbf{u}} L_{\mathbf{u}} + small$$

There are 2 goals in choosing the update function to be $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta(\mathbf{w}_t \cdots \mathbf{x}_t - y_t) \cdot \mathbf{x}_t$: (1) Want loss of $\mathbf{w}_{t+1}$ on $\mathbf{x}_t$, y_t to be small. This means we want to minimize $(\mathbf{w}_{t+1} \cdot \mathbf{x}_t - y_t)^2$ (2) Want $\mathbf{w}_{t+1}$ close to $\mathbf{w}_t$ so that we do not forget everything we learnt so far. And this means we want to minimize $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2$.

Therefore to sum up, we want to minimize

$$\eta(\mathbf{w}_{t+1} \cdot \mathbf{x}_t - y_t)^2 + \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2$$

If we take the derivative of the above equation and set it to zero, we have

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta(\mathbf{w}_{t+1} \cdot \mathbf{x}_t - y_t) \cdot \mathbf{x}_t$$

Instead of solving $\mathbf{w}_{t+1}$, we approximate the term $\mathbf{w}_{t+1}$ inside the parenthesis and change it to $\mathbf{w}_t$. The reason we can do this is because $\mathbf{w}_{t+1}$ does not change much from $\mathbf{w}_t$. Therefore we have

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta(\mathbf{w}_t \cdot \mathbf{x}_t - y_t) \cdot \mathbf{x}_t$$

which is the update function stated in the algorithm.

Now let's state a theorem:

Theorem 1.1 If we assume on every round t, $\|\mathbf{x}_t\|_2 \leq 1$, then:

$$L_{WH} \leq \min_{\mathbf{u} \in \mathbb{R}^n} [\frac{L_{\mathbf{u}}}{1-\eta} + \frac{\|\mathbf{u}\|_2^2}{\eta}]$$

From this theorem, we have $\forall \mathbf{u}$:

$$L_{WH} \leq \frac{1}{1-\eta} \cdot L_{\mathbf{u}} + \frac{\|\mathbf{u}\|_2^2}{\eta}$$

If we divide T on both side, we have:

$$\frac{L_{WH}}{T} \leq \frac{1}{1-\eta} \cdot \frac{L_{\mathbf{u}}}{T} + \frac{\|\mathbf{u}\|_2^2}{\eta T}$$

The term $\frac{\|\mathbf{u}\|_2^2}{\eta T}$ goes to 0 when T gets large. And we can choose η small enough to make $\frac{1}{1-\eta}$ to be close to 1. Therefore we have the rate that the algorithm is suffering loss is close to rate that $L_{\mathbf{u}}$ is suffering loss.

Now let's prove the theorem:

Proof: Pick any $\mathbf{u} \in \mathbb{R}^n$. First let's define some terms:

$$\begin{aligned}
\Phi_t &= \|\mathbf{w}_t - \mathbf{u}\|_2^2 && \text{(measure of progess)} \\
l_t &= \mathbf{w}_t \cdot \mathbf{x}_t - y_t = \hat{y}_t - y_t && \text{(notice } l_t^2 \text{ is the loss of WH on round } t\text{)} \\
g_t &= \mathbf{u} \cdot \mathbf{x}_t - y_t && (g_t^2 \text{ is the loss of } \mathbf{u} \text{ on round } t\text{)} \\
\mathbf{\Delta}_t &= \eta(\mathbf{w}_t \cdot \mathbf{x}_t - y_t) \cdot \mathbf{x}_t = \eta l_t \mathbf{x}_t \\
\mathbf{w}_{t+1} &= \mathbf{w}_t - \mathbf{\Delta}_t
\end{aligned}$$

Our main claim is that the change of potential is:

$$\Phi_{t+1} - \Phi_t \leq -\eta l_t^2 + \frac{\eta}{1-\eta} \cdot g_t^2 \tag{1}$$

This shows that l_t^2 tends to drive potential down while g_t^2 tends to drive potential up.

Now assume (1) holds. Notice that total change in potential should be non-negative. And also we initialize $\mathbf{w}_1 = \mathbf{0}$. So we have the following inequality:

$$\begin{aligned}
-\|\mathbf{u}\|_2^2 = -\Phi_1 &\leq \Phi_{T+1} - \Phi_1 \\
&= \Phi_{t+1} - \Phi_t + \Phi_t - \Phi_{t-1} + \cdots + \Phi_2 - \Phi_1 \\
&= \sum_{t=1}^{T} (\Phi_{t+1} - \Phi_t) \\
&\leq \sum_{t=1}^{T} [-\eta l_t^2 + \frac{\eta}{1-\eta} g_t^2] \\
&= -\eta \sum_t l_t^2 + \frac{\eta}{1-\eta} \sum_t g_t^2 \\
&= -\eta L_{WH} + \frac{\eta}{1-\eta} L_{\mathbf{u}}
\end{aligned}$$

Now we solve for L_{WH}, we get

$$L_{WH} \leq \frac{1}{1-\eta} L_{\mathbf{u}} + \frac{\|\mathbf{u}\|_2^2}{\eta}$$

And since this inequality holds for all $\mathbf{u}$, we have:

$$L_{WH} \leq \min_{\mathbf{u}\in\mathbb{R}}[\frac{1}{1-\eta}L_{\mathbf{u}} + \frac{\|\mathbf{u}\|_2^2}{\eta}]$$

which is the theorem.

Now let's go back to prove (1):

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= \|\mathbf{w}_{t+1} - \mathbf{u}\|^2 - \|\mathbf{w}_t - \mathbf{u}\|^2 \\
&= \|\mathbf{w}_t - \mathbf{u} - \boldsymbol{\Delta}_t\|^2 - \|\mathbf{w}_t - \mathbf{u}\|^2 \\
&= \|\Delta_t\|^2 - 2(\mathbf{w}_t - \mathbf{u}) \cdot \boldsymbol{\Delta}_t + \|\mathbf{w}_t - \mathbf{u}\|^2 - \|\mathbf{w}_t - \mathbf{u}\|^2 \\
&= \|\Delta_t\|^2 - 2(\mathbf{w}_t - \mathbf{u}) \cdot \boldsymbol{\Delta}_t \\
&= \|\Delta\|^2 - 2(\mathbf{w} - \mathbf{u}) \cdot \boldsymbol{\Delta} \quad \text{(dropping subscript } t \text{ since it doesn't affect the proof)} \\
&= \eta^2 l^2 \|\mathbf{x}\|^2 - 2\eta l \mathbf{x} \cdot (\mathbf{w} - \mathbf{u}) \\
&= \eta^2 l^2 \|\mathbf{x}\|^2 - 2\eta l (\mathbf{w} \cdot \mathbf{x} - \mathbf{u} \cdot \mathbf{x} - y + y) \\
&= \eta^2 l^2 \|\mathbf{x}\|^2 - 2\eta l [(\mathbf{w} \cdot \mathbf{x} - y) - (\mathbf{u} \cdot \mathbf{x} - y)] \\
&= \eta^2 l^2 \|\mathbf{x}\|^2 - 2\eta l [l - g] \\
&= \eta^2 l^2 \|\mathbf{x}\|^2 - 2\eta l^2 + 2\eta l g \\
&\leq \eta^2 l^2 - 2\eta l^2 + 2\eta l g && (\|\mathbf{x}\|^2 \leq 1) \\
&\leq (\eta^2 - 2\eta) l^2 + \frac{2\eta[\frac{g^2}{1-\eta} + l^2(1-\eta)]}{2} && (ab \leq \tfrac{a^2+b^2}{2}) \\
&= (\eta^2 - 2\eta) l^2 + \eta[\frac{g^2}{1-\eta} + l^2(1-\eta)] \\
&= -\eta l^2 + \frac{\eta}{1-\eta} g^2
\end{aligned}$$

2 Families of Online Algorithm

The two goals of the learning algorithm are minimizing the loss of $\mathbf{w}_{t+1}$ on $\mathbf{x}_t$ and y_t, and minimizing the distance between $\mathbf{w}_{t+1}$ and $\mathbf{w}_t$. So to generalize, we are trying to minimize

$$\eta L(\mathbf{w}_{t+1}, \mathbf{x}_t, y_t) + d(\mathbf{w}_{t+1}, \mathbf{w}_t)$$

So if we use the Euclidean norm as our distance measurement, then the above function becomes:

$$\eta L(\mathbf{w}_{t+1}, \mathbf{x}_t, y_t) + \|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$

So if we try to optimize the above function, we have the update equation:

$$\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t - \eta \nabla_{\mathbf{w}} L(\mathbf{w}_{t+1}, \mathbf{x}_t, y_t) \\
&\approx \mathbf{w}_t - \eta \nabla_{\mathbf{w}} L(\mathbf{w}_t, \mathbf{x}_t, y_t)
\end{aligned}$$

Notice that we use $\mathbf{w}_t$ to approximate $\mathbf{w}_{t+1}$ when we calculate $\mathbf{w}_{t+1}$. This is called the Gradient Descent Algorithm.

Alternatively, we can use relative entropy as a measure of distance. So $d(\mathbf{w}_t, \mathbf{w}_{t+1}) = RE(\mathbf{w}_t \| \mathbf{w}_{t+1})$. Now we can have the update function as

$$w_{t+1,i} = \frac{w_{t,i} \cdot \exp(\eta \frac{\partial L(\mathbf{w}_{t+1}, \mathbf{x}_t, y_t))}{\partial w_i})}{\mathcal{Z}_t}$$

This is called the Exponentiated Gradient Algorithm, or "EG" algorithm. We need to change the norm: $\|\mathbf{x}_t\|_\infty \leq 1$ and $\|\mathbf{u}\|_1 = 1$. It's also possible to prove a bound on this update equation, but we skip it in this class.

3 Online Algorithm in a Batch Setting

We can modify the online algorithms slightly so that we can use them in the batch learning settings. Let's take a look at one example in a linear regression setting. In a linear regression setting, training and test samples are drawn i.i.d from a fixed distribution $\mathcal{D}$. So we have $\mathcal{S} = \langle(\mathbf{x}_1, y_1) \dots (\mathbf{x}_m, y_m)\rangle$ where $(x_i, y_i) \sim \mathcal{D}$. Our goal is to find $\mathbf{v}$ with low risk, where risk is defined to be

$$R_{\mathbf{v}} = E_{(\mathbf{x},y)\sim\mathcal{D}}[(\mathbf{v} \cdot \mathbf{x} - y)^2]$$

We want to find $\mathbf{v}$ such that $R_{\mathbf{v}}$ is small compared to $\min_{\mathbf{u}} R_{\mathbf{u}}$.

Now we can apply WH algorithm to the data as follows:
(1) run WH on $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$, and calculate $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m$.

(2) Combine the vectors:

$$\mathbf{v} = \frac{1}{m} \sum_{t=1}^{m} \mathbf{w}_t$$

and output $\mathbf{v}$. We choose to output the average of all the $\mathbf{w}_t$'s because we can prove something theoretically good about it, which is not necessarily the case for the last vector $\mathbf{w}_m$.

Now let's state another theorem:

Theorem 3.1

$$E_{\mathcal{S}}[R_{\mathbf{v}}] \leq \min_{\mathbf{u}\in\mathbb{R}^n} [\frac{R_{\mathbf{u}}}{1-\eta} + \frac{\|\mathbf{u}\|^2}{\eta m}]$$

If we divide T on both side of the equation above and if η is chosen to be small, we can see that $\frac{R_{\mathbf{v}}}{T}$ will be close to $\frac{R_{\mathbf{u}}}{T}$ when T is large. **Proof**:

There are three observations needed in the proof:

(1):
Let $\mathbf{x}, y$ be a random test example from $\mathcal{D}$. Then we have

$$(\mathbf{v} \cdot \mathbf{x} - y)^2 \leq \frac{1}{m} \sum_{t=1}^{m} (\mathbf{w}_t \cdot \mathbf{x}_t - y)^2$$

Proof for (1):

$$
\begin{aligned}
(\mathbf{v}\cdot\mathbf{x}-y)^2 &= [(\frac{1}{m}\sum_{t=1}^{m}\mathbf{w}_t)\cdot\mathbf{x}-y]^2 \\
&= [(\frac{1}{m}\sum_{t=1}^{m}\mathbf{w}_t\cdot\mathbf{x})-y]^2 \\
&= [\frac{1}{m}\sum_{t=1}^{m}(\mathbf{w}_t\cdot\mathbf{x}-y)]^2 \\
&\le \frac{1}{m}\sum_{t}(\mathbf{w}_t\cdot\mathbf{x}-y)^2 && \text{(convexity of } f(x)=x^2\text{)}
\end{aligned}
$$

(2):

$$E[(\mathbf{u}\cdot\mathbf{x}_t-y_t)^2]=E[(\mathbf{u}\cdot\mathbf{x}-y)^2]$$

The above expectation is with respect to the random choice of $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_m,y_m)$ and $(\mathbf{x},y)$. This is because $(\mathbf{x}_t,y_t)$and $(\mathbf{x},y)$ are from the same distribution.

(3):

$$E[(\mathbf{w}_t\cdot\mathbf{x}_t-y_t)^2]=E[(\mathbf{w}_t\cdot\mathbf{x}-y)^2]$$

This is because $\mathbf{w}_t$ only depends on the first $t-1$ samples but doesn't depend on $(\mathbf{x}_t,y_t)$.

Now let's start the proof:

$$
\begin{aligned}
E_{\mathcal{S}}[R_{\mathbf{v}}] &= E_{\mathcal{S},(\mathbf{x},y)}[(\mathbf{v}\cdot\mathbf{x}-y)^2] \\
&\le E[\frac{1}{m}\sum_{t}(\mathbf{w}_t\cdot\mathbf{x}-y)^2] && \text{(using observation (1))} \\
&= \frac{1}{m}\sum_{t}E[(\mathbf{w}_t\cdot\mathbf{x}-y)^2] \\
&= \frac{1}{m}\sum_{t}E[(\mathbf{w}_t\cdot\mathbf{x}_t-y_t)^2] && \text{(observation (3))} \\
&= \frac{1}{m}E[\sum_{t}(\mathbf{w}_t\cdot\mathbf{x}_t-y_t)^2] \\
&\le \frac{1}{m}E[\frac{\sum_t(\mathbf{u}\cdot\mathbf{x}_t-y_t)^2}{1-\eta}+\frac{\|\mathbf{u}\|^2}{\eta}] && \text{(by WH bound)} \\
&= \frac{1}{m}[\frac{\sum_t E[(\mathbf{u}\cdot\mathbf{x}_t-y_t)^2]}{1-\eta}+\frac{\|\mathbf{u}\|^2}{\eta}] \\
&= \frac{1}{m}[\frac{\sum_t E[(\mathbf{u}\cdot\mathbf{x}-y)^2]}{1-\eta}]+\frac{\|\mathbf{u}\|^2}{\eta m} && \text{(by observation (2))} \\
&= \frac{R_{\mathbf{u}}}{1-\eta}+\frac{\|\mathbf{u}\|^2}{\eta m}
\end{aligned}
$$

and we have completed the proof.