Foundations of Machine Learning — Homework Assignment 3

Anirudhan J Rajagopalan
N18824115
ajr619

December 09, 2015

C. Randomized Halving

1. Psuedo code

Algorithm 1 Randomized Halving

```
$H_1 \leftarrow H$
for t ← 1 to T do
    $RECEIVE(x_t)$
    $r_t \leftarrow \frac{\sum_{i:y_{t,i}=1} 1}{|H_t|}$
    $p_t \leftarrow 1$
    if $r_t \leq \frac{3}{4}$ then
        $p_t \leftarrow [\frac{1}{2}\log_2 \frac{1}{1-r_t}]$
    $\hat{y_t} \leftarrow GetRandomNumberWithProbability([1,0],[p_t, 1-p_t])$
    $RECEIVE(y_t)$
    if $\hat{y_t} \neq y_t$ then
        $H_{t+1} \leftarrow \{c \in H_t : c(x_t) = y_t\}$
return $H_{T+1}$
```

2. Prove $\forall t \geq 1, E[\mu] \leq \frac{\phi_t - \phi_{t+1}}{2}$

Given: Potential function: $\phi_t = \log_2 \mid H_t \mid$ and $\mu_t = 1_{y_t \neq \hat{y_t}}$

Proof:

We are only considering the case when the predicted value $\hat{y_t}$ is not equal to the received value y_t. Assume that our prediction is 1 but the actual result is zero. So the probability of predicting 1 is the probability of mistake. In this case: The value of expectation can be written as

$$\begin{aligned} E[\mu_t] =& p_t * 1 + (1 - p_t) * 0 \\ =& p_t * 1 \end{aligned}$$

The probability of predicting 1 by the randomized algorithm is the probability of making a mistake since we are only considering the cases in which we make mistakes ($\mu_t = 1_{y_t \neq \hat{y_t}}$)

Therefore,

$$\begin{aligned} E[\mu_t] =& p_t \\ =& [\frac{1}{2}\log_2 \frac{1}{1-r_t}] 1_{r_t \leq \frac{3}{4}} + 1_{r_t > \frac{3}{4}} \\ \leq& [\frac{1}{2}\log_2 \frac{1}{1-r_t}] \end{aligned}$$

Since, the Expectation will be 1 when $r_t > \frac{3}{4}$ which corresponds to the maximum expectation here, we can upper bound the expectation by using $r_t \leq \frac{3}{4}$ as $[\frac{1}{2}\log_2 \frac{1}{1-r_t}]$ equals 1 when $r_t = \frac{3}{4}$

Let E_1, E_0, E_t denote the number of experts predicting 1, 0 and the total number of experts in a round t.

$$\begin{aligned}
E[\mu_t] \leq& [\frac{1}{2}\log_2 \frac{1}{1-r_t}] \\
=& [\frac{\log_2 \frac{1}{1-r_t}}{2}] \\
=& [\frac{\log_2 \frac{1}{1-r_t}}{2}] \\
=& [\frac{\log_2 \frac{1}{1-\frac{|E_1|}{|H_t|}}}{2}] \\
=& [\frac{\log_2 \frac{|H_t|}{|H_t|-|E_1|}}{2}] \\
=& [\frac{\log_2 \frac{|H_t|}{|E_0|}}{2}] \\
=& [\frac{\log_2 \mid H_t \mid - \log_2 \mid E_0 \mid}{2}] \\
=& [\frac{\log_2 \mid H_t \mid - \log_2 \mid H_{t+1} \mid}{2}] \\
=& [\frac{\phi_t - \phi_{t+1}}{2}] \\
\therefore E[\mu_t] \leq& \frac{\phi_t - \phi_{t+1}}{2}
\end{aligned}$$

Similarly we can show that when our algorithm predicts zero and when the actual label is 1 we can show that $E[\mu_t] \leq \frac{\phi_t - \phi_{t+1}}{2}$

3. Expected number of mistakes.

Given: N be the total number of experts at the beginnning of the iterations (denoted by H_1). Since we are considering a relizable scenario at the end of the algorithm the number of experts should be atleast one. So $H_T = 1$.

To Prove: The expected number of mistakes made by Randomized Halving is at most $\frac{1}{2}\log_2 N$

Proof:

Lets consider the total expectation of mistakes of the Randomized Halving algorithm over T iterations.

$$\begin{aligned}
E[\mu_T] \leq& \sum_{t=1}^{T} \frac{\phi_t - \phi_{t+1}}{2} \\
\leq& \frac{(\phi_{H_1} - \phi_{H_2}) + (\phi_{H_2} - \phi_{H_3}) \cdots (\phi_{H_{T-2}} - \phi_{H_{T-1}}) + (\phi_{H_{T-1}} - \phi_{H_T})}{2} \\
\leq& \frac{\phi_{H_1} - \phi_{H_T}}{2} \\
\leq& \frac{\phi_N - \phi_T}{2} \\
\leq& \frac{\phi_N}{2} \\
\leq& \frac{1}{2}\log_2 N
\end{aligned}$$

Hence proved.

Here $\phi_T = 0$ because the number of experts at line T is 1. Therefore, log of T will be zero.

4. [Bonus Question]

As we have seen in the previous answer, the mistakes made by the randomized algorithm is bounded by $\frac{1}{2}\log_2 N$. This upper bound is dependent only on the number of initial experts, N. Therefore any randomized algorithm that is dependent on the opinion of the experts to generate its predictions will have similar upper bound of $\lfloor \frac{1}{2}\log_2 N \rfloor$. The floor function is used as mistakes are natural numbers.

A. Boosting-type Algorithm

1. Bound of $1_{u\leq 0}$ Proof of convexity and differentiability

Given: $\phi_p(u) = \max((1+u)^p, 0)$

To prove: 1. Function $\phi_p(u)$ is convex and differentiable
And 2. $\forall u \in \mathbb{R}$ and $p > 1, 1_{u\leq 0} \leq \phi_p(-u)$

Proof of $\forall u \in \mathbb{R}$ and $p > 1, 1_{u\leq 0} \leq \phi_p(-u)$ There are three cases here:

When $u = 0$

$$\begin{aligned}
1_{u\leq 0} =& 1 \\
\phi_p(-u) =& \max((1-0)^p, 0) = 1 = 1_{u\leq 0}
\end{aligned}$$

When $u < 0$

$$\begin{aligned}1_{u \leq 0} =& 1 \\ \phi_p(-u) =& \max((1-u)^p, 0) \\ =& \max((1+u)^p, 0) \text{Since u is negative, -u is positive} \\ >& 1 \\ >& 1_{u \leq 0}\end{aligned}$$

When $u > 0$

$$\begin{aligned}1_{u \leq 0} =& 0 \\ \phi_p(-u) =& \max((1-u)^p, 0) \\ =& \begin{cases} 0, & \text{if p is odd} \\ (1-u)^p, & \text{if p is positive} \end{cases} \\ \geq& 0 \quad \forall \quad \mathbb{R} \geq 1_{u \leq 0}\end{aligned}$$

Hence proved

Proof of Convexity and Differentiability: The function can be written as a piecewise function based on the value of p:

When p is even (2 and higher value even numbers)

$$\phi_p(u) = (1+u)^p$$

Since p > 1, the function is differentiable.

$$\begin{aligned}\phi_p^{'}(u) =& p(1+u)^{p-1} \\ \phi_p^{''}(u) =& (p-1)p(1+u)^{p-2} \\ \because p > 1, \phi_p^{''}(u) >& 0\end{aligned}$$

Since the double derivative is greater than zero, the function is convex and differentiable.

When p is odd (1 and higher value odd numbers) the function can be defined using the piecewise function

$$\phi_p\,(x) = \begin{cases} 0, & \text{if } u \leq -1 \\ (1+u)^p, & otherwise \end{cases}$$

The function is piecewise continuous and differentiable. We have to check that the function is differentiable at u = -1 to show that the function is differentiable and continuous. We should also check the double derivative to show that the function is convex.

The first derivative can be found using limits.

$$\begin{aligned}
\lim_{h \to 0+} \frac{\phi_p(-1+h) - \phi_p(-1)}{h} &= \lim_{h \to 0+} \frac{(1 + (-1+h))^p - (1-1)^p}{h} \\
&= \lim_{h \to 0+} \frac{h^p - 0}{h} \\
&= \lim_{h \to 0+} h^{p-1} \\
&=0
\end{aligned}$$

Also

$$\begin{aligned}
\lim_{h \to 0-} \frac{\phi_p(-1+h) - \phi_p(-1)}{h} &= \lim_{h \to 0+} \frac{(1 + (-1+h))^p - (1-1)^p}{h} \\
&= \lim_{h \to 0-} \frac{h^p - 0}{h} \\
&= \lim_{h \to 0-} h^{p-1} \\
&=0
\end{aligned}$$

Since both the limit values are equal, the function is differentiable. Therefore:

$$\phi_p^{'}(x) = \begin{cases} 0, & \text{if } u \leq -1 \\ p(1+u)^{p-1}, & otherwise \end{cases}$$

For showing the function is convex, we find the double differeniable using limits for $\phi_p^{'}$ and show that it is non negative

$$\begin{aligned}
\lim_{h \to 0+} \frac{\phi_p^{'}(-1+h) - \phi_p^{'}(-1)}{h} &= \lim_{h \to 0+} \frac{p(1 + (-1+h))^{p-1} - 0}{h} \\
&= \lim_{h \to 0+} \frac{p(h)^{p-1}}{h} \\
&= \lim_{h \to 0+} p(h)^{p-2} \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\lim_{h\to 0-} \frac{\phi_p^{'}(-1+h) - \phi_p^{'}(-1)}{h} &= \lim_{h\to 0-} \frac{p(1+(-1+h))^{p-1} - 0}{h} \\
&= \lim_{h\to 0-} \frac{p(h)^{p-1}}{h} \\
&= \lim_{h\to 0-} p(h)^{p-2} \\
&= 0
\end{aligned}$$

Since the left and right derivatives are equal, the function is $\phi_p^{'}$ is differentiable. Which implies that the original function ϕ_p is double differentiable.

$$\phi_p^{''}(x) = \begin{cases} 0, & \text{if } u \le -1 \\ (p-1)p(1+u)^{p-2}, & otherwise \end{cases}$$

Since in both cases the double differentiable is non negative, the function is convex and differentiable.

2. Derive boosting type algorithm

We can build an algorithm similar to adaboost using coordinate descent. The algorithm and the explanation are given below.

Algorithm 2 Boosting type algorithm

```
$S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$
for i ← 1 to m do
    $D_1(i) \leftarrow \frac{1}{m}$
for i ← 1 to T do
    $h_t \leftarrow$ base classifier in H with small error $\epsilon_t = \Pr_{D_t}[h_t(x_i) \neq y_i]$
    $\alpha_t \leftarrow \eta$ obtained by searching non-linear equation of $\eta$
    $f_t \leftarrow f_{t-1} + \alpha_t h_t$
for i ← 1 to m do
    $D_t(i) \leftarrow \frac{1}{m} \frac{\phi_p^{'}(-y_i f_t(x_i))}{\sum_{i=1}^{m} \phi_p^{'}(-y_i f_t(x_i))}$
return $h = sgn(f_T)$
```

The objective function is defined as

$$F(\alpha) = \sum_{i=1}^{m} \phi_p(-y_i f_T(x_i))$$

Where

$$F_T(x_i) = \sum_{t=1}^{T} \alpha_t h_t(x_i)$$

and the weighted distribution of the sample is

$$D_t(i) \leftarrow \frac{1}{m} \frac{\phi_p^{'}(-y_i f_t(x_i))}{\sum_{i=1}^{m} \phi_p^{\prime}(-y_i f_t(x_i))}$$

The weighted distribution is initialized to $\frac{1}{m}$ initially.

Let e_t denote the unit vector corresponding to the coordinate in $\mathbb{R}^n$ and let α_{t-1} denote the vector based on the (t-1) coefficients.

$$\alpha_{t-1} = \begin{cases} (\alpha_1, \alpha_2, \ldots, \alpha_{t-1}, 0, 0, \ldots, 0)^T & \text{if(t} - 1) > 0 \\ 0 & otherwise \end{cases}$$

Direction

Coordinate descent selects the direction e_t that minimizes the directional derivative.

$$e_t = argmin_t [\frac{dF(\alpha_{t-1} + \eta e_t)}{d\eta}]_{\eta=0}$$

where

$$\begin{aligned} F(\alpha_{t-1} + \eta e_t) &= \sum_{i=1}^{m} \phi_p \left(-y_i \sum_{j=1}^{t-1} (\alpha_j + \eta e_t) h_j \right) \\ &= \sum_{i=1}^{m} \phi_p \left(-y_i \sum_{j=1}^{t-1} \alpha_j h_j(x_i) - \eta y_i h_t(x_i) \right) \end{aligned}$$

Also,

$$\begin{aligned}
\frac{dF(\alpha_{t-1}+\eta e_t)}{d\eta} =& -\sum_{i=1}^{m} y_i h_t(x_i).\phi_p^{'}\left(-y_i\sum_{j=1}^{t-1}\alpha_j h_j(x_i) - \eta y_i h_t(x_i)\right) \\
\left[\frac{dF(\alpha_{t-1}+\eta e_t)}{d\eta}\right]_{\eta=0} =& -\sum_{i=1}^{m} y_i h_t(x_i).\phi_p^{'}\left(-y_i\sum_{j=1}^{t-1}\alpha_j h_j(x_i)\right) \\
=& \sum_{i=1}^{m} y_i h_t(x_i).D(i).\left(m\sum_{k=1}^{m}\phi_p^{'}(-y_k f_t(x_k))\right) \\
=& -\left(\sum_{y_i h_t=-1} D(i) + \sum_{y_i h_t=1} D(i)\right)\left(m\sum_{k=1}^{m}\phi_p^{'}(-y_k f_t(x_k))\right) \\
=& -((1-\epsilon_t-\epsilon_t))\left(m\sum_{i=1}^{m}\phi_p^{'}(-y_i f_t(x_i))\right) \\
=& (2\epsilon_t - 1)\left(m\sum_{i=1}^{m}\phi_p^{'}(-y_i f_t(x_i))\right) \\
\propto& (2\epsilon_t - 1)
\end{aligned}$$

Because of the above equation, it can be seen that the hypothesis chosen by the algorithm minimizes the mis-classification error.

Step

We can obtain the step by minimizing the derivative of dF with respect to η to find the direction along e_t

$$-\sum_{i=1}^{m} y_i h_t(x_i).\phi_p^{'}\left(-y_i\sum_{j=1}^{t-1}\alpha_j h_j(x_i) - \eta y_i h_t(x_i)\right) = 0$$

Since $\phi_p^{'}$ is defined as

$$= \begin{cases} p(1+u)^{p-1} & \text{if}(1+u)^p \geq 0 \\ 0 & otherwise \end{cases}$$

Let us define a set $I = \left\{i \in [1,m] : \phi_p^{'}\left(-y_i \sum_{j=1}^{t-1}\alpha_j h_j(x_i) - \eta y_i h_t(x_i)\right) \neq 0\right\}$ Since we only have to consider the values of i in which the derivative is non zero. Now, by replacing $\phi_p^{'}(-u)$ with $p(1-u)^{p-1}$ we have

$$\begin{aligned}
-p\sum_{i\in I} y_i h_t(x_i)D(i)\left(m\sum_{k=1}^{m}\phi_p^{'}(y_k f_{t-1}(x_k)-\eta y_k h_t(x_k))\right) &=0\\
-\sum_{i\in I} y_i h_t(x_i)D(i)\left(mp\sum_{k\in I}\left(1-y_i\sum_{j=1}^{t-1}\alpha_j h_j(x_i)-\eta y_i h+t(x_i)\right)^{p-1}\right) &=0\\
\sum_{y_i h_t=+1}(1-\epsilon_t)\sum_{k\in I}\left(1-y_i\sum_{j=1}^{t-1}\alpha_j h_j(x_i)-\eta\right)^{p-1}-\epsilon_t\sum_{k\in I}\left(1-y_i\sum_{j=1}^{t-1}\alpha_j h_j(x_i)-\eta\right)^{p-1} &=0
\end{aligned}$$

by removing the constants and negative signs. This gives us a (p-1) monic polynomial in η:

$$\eta^{p-1}+a_{p-2}(\epsilon_t)\eta^{p-2}+\cdots+a_1(\epsilon_t)\eta+a_0(\epsilon_t)=0$$

This can be solved using Newtons' method to find the value of step size η

Generalization Bound

A family of functions (H) taking values in +1, -1 with VC-dimension d for any $\delta > 0$ with probability atleast $1-\delta$, it holds for any $h \in conv(H)$:

$$R(h) \leq \hat{R_p}(h) + \frac{2}{p}\sqrt{\frac{2d\log\frac{em}{d}}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

The above expression is true for all ensemble methods (Refer Corollary 6.2, page 133 of Foundations of Machine Learning, Mehryar Mohri et al.) it will be true for the algorithm described above. Also we cannot have an emperical bound on $\hat{R}_p(h)$ in terms of misclassification error as there is not closed form expression of α_t

L2 regularized Maxent

1

To Prove: $\forall \delta > 0$ with probability $1 - \delta$ the following inequality holds

$$\|E_{x\sim D}[\Phi(x)] - E_{x\sim S}[\Phi(x)]\|_2 \leq \sqrt{\frac{2r^2}{m}}\left(1 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right)$$

Proof: Let $\beta = \|E_{x\sim D}[\Phi(x)] - E_{x\sim S}[\Phi(x)]\|_2$. Also let $|\beta - \beta'|$ be the quantity in which we we modify β by one dimension of x_t. $\beta' = \beta(x_1, x_2, \ldots, x_i', \ldots, x_m)$ Then,

$$\begin{aligned}|\beta - \beta'| &= \|E_{x\sim D}[\Phi(x)] - E_{x\sim S}[\Phi(x)] - E'_{x\sim D}[\Phi(x)] + E'_{x\sim S}[\Phi(x)]\|_2 \\ &= \left\|E_{x\sim D}[E_{x\sim S}[\Phi(x)]] - \frac{1}{m}\sum_{i=1}^{m}\Phi(x_i) - E'_{x\sim D}[E'_{x\sim S}[\Phi(x)]] + \frac{1}{m}\sum_{i=1}^{m-1}(\Phi(x_i) + \Phi(x_m'))\right\|_2 \\ &= \left\|\frac{\Phi(x_m') - \Phi(x_m)}{m} + E_{x\sim D}[E_{x\sim S}[\Phi(x)]] - E'_{x\sim D}[E'_{x\sim S}[\Phi(x)]]\right\|_2\end{aligned}$$

Since $\Phi(x_i) \leq r$;

$$\begin{aligned}\|(\Phi(x_m') + E_{x\sim D}[\Phi(x_i)]) - (\Phi(x_m) + E_{x\sim D}[\Phi'(x_i)])\| &\leq 2r \\ |\beta - \beta'| &\leq \frac{2r}{m}\end{aligned}$$

Using McDiarmid's Inequality,

$$\begin{aligned}\Pr[\beta - E[\beta] \geq \epsilon] &\leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m}(2r/m^2)}\right) \\ &= \exp\left(-\frac{2\epsilon^2}{(2r/m)}\right)\end{aligned}$$

Using δ as exponential term:

$$\begin{aligned}
\delta =& \exp\left(-\frac{2\epsilon^2}{(2r/m)}\right) \Rightarrow \epsilon = \sqrt{\frac{2r^2}{m}\log\left(\frac{1}{\delta}\right)} \\
\beta - E[\beta] \geq& \epsilon \Rightarrow \beta - E[\beta] \geq \sqrt{\frac{2r^2}{m}\log\left(\frac{1}{\delta}\right)} \\
\beta - E[\beta] \leq& \sqrt{\frac{2r^2}{m}\log\left(\frac{1}{\delta}\right)} \\
\beta \leq& E[\beta] + \sqrt{\frac{2r^2}{m}\log\left(\frac{1}{\delta}\right)}
\end{aligned}$$

By using Jensen's inequality:

$$\begin{aligned}
E[\beta] =& E\left[\sqrt{\beta^2}\right] \\
\leq& \sqrt{E\left[\beta^2\right]}
\end{aligned}$$

Also:

$$\begin{aligned}
\sigma^2[\beta] =& E\left[\beta^2\right] - E[\beta]^2 \\
\therefore E[\beta] \leq& \sqrt{E\left[\beta^2\right]} \\
=& \sqrt{\sigma^2[\beta] + E[\beta]^2}
\end{aligned}$$

Since,

$$\Phi(x) \leq r \Rightarrow E[||E_{x\sim D}\left[\Phi(x)\right] - E_{x\sim S}\left[\Phi(x)\right]||]^2 \leq \frac{r^2}{m}$$

Thus we have,

$$\begin{aligned}
\beta \leq& E[\beta] + \sqrt{\frac{2r^2}{m}\log\left(\frac{1}{\delta}\right)} \\
=& \sqrt{\sigma^2[\beta] + E[\beta]^2} + \sqrt{\frac{2r^2}{m}\log\left(\frac{1}{\delta}\right)}
\end{aligned}$$

Which implies:

$$\left\|E_{x\sim D}\left[\Phi(x)\right] - E_{x\sim S}\left[\Phi(x)\right]\right\|_2 \leq \sqrt{\frac{2r^2}{m}}\left(1 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right).$$

2

To Prove: $\left\|\hat{w} - w_D\right\|_2 \leq \frac{1}{\lambda}\left\|E_{x\sim S}\left[\Phi\left(x\right)\right] - E_{x\sim D}\left[\Phi\left(x\right)\right]\right\|_2$

Proof: We know that when $\hat{w} \neq w_D$:

$$\begin{aligned} L_Q(w) =& E_{x\sim Q}\left[-\log\left(p_w\left(x\right)\right)\right] \\ =& -E_{x\sim Q}\left[\log\left(\frac{\exp\left(w.\Phi\left(x\right)\right)}{Z}\right)\right] \\ =& -E_{x\sim Q}\left[w.\Phi\left(x\right) - \log\left(Z\right)\right] \\ =& \log\left(Z\right) - w.E_{x\sim Q}\left[\Phi\left(x\right)\right] \\ =& q\left(w\right) - w.E_{x\sim Q}\left[\Phi\left(x\right)\right] \end{aligned}$$

by substituting $q(w) = \log(Z)$

$$\begin{aligned} J_S(w) =& \frac{\lambda}{2}\left\|w\right\|_2^2 + L_S(w) \\ =& \frac{\lambda}{2}\left\|w\right\|_2^2 + q\left(w\right) - w.E_{x\sim S}\left[\Phi\left(x\right)\right] \\ J_D(w) =& \frac{\lambda}{2}\left\|w\right\|_2^2 + L_D(w) \\ =& \frac{\lambda}{2}\left\|w\right\|_2^2 + q\left(w\right) - w.E_{x\sim D}\left[\Phi\left(x\right)\right] \end{aligned}$$

Taking first derivative and equating to zero:

$$\begin{aligned} \left[\frac{dJ_S(w)}{dw}\right] w = \hat{w} = 0 \Rightarrow \lambda\hat{w} + \nabla q(\hat{w}) - E_{x\sim S}\left[\Phi\left(x\right)\right] =& 0 \\ \left[\frac{dJ_D(w)}{dw}\right]_{w=w_D} = 0 \Rightarrow \lambda w_D + \nabla q(w_D) - E_{x\sim D}\left[\Phi\left(x\right)\right] =& 0 \end{aligned}$$

From the above equations:

$$\begin{aligned} \lambda\left(w_D - \hat{w}\right) =& \left(E_{x\sim D}\left[\Phi\left(x\right)\right] - E_{x\sim S}\left[\Phi\left(x\right)\right]\right) - \left(\nabla q(w_D) - \nabla q(\hat{w})\right) \\ \lambda\left\|w_D - \hat{w}\right\|_2^2 =& \left(w_D - \hat{w}\right)\cdot\left(E_{x\sim D}\left[\Phi\left(x\right)\right] - E_{x\sim S}\left[\Phi\left(x\right)\right]\right) - \left(w_D - \hat{w}\right)\cdot\left(\nabla q(w_D) - \nabla q(\hat{w})\right) \end{aligned}$$

Since Z (sum of components) is a convex function of w, log (z) is also a convex function. Using the identity

$$(\nabla q(w_1) - \nabla q(w_2)) \cdot (w_1 - w_2) \geq 0$$

we have:

$$\begin{aligned}\lambda \left\| w_D - \hat{w} \right\|_2^2 &\leq (w_D - \hat{w}) \cdot (E_{x\sim D}\left[\Phi\left(x\right)\right] - E_{x\sim S}\left[\Phi\left(x\right)\right]) \\ &= (\hat{w} - w_D) \cdot (E_{x\sim S}\left[\Phi\left(x\right)\right] - E_{x\sim D}\left[\Phi\left(x\right)\right]) \\ \lambda \left\| \hat{w} - w_D \right\|_2^2 &\leq \left\| \hat{w} - w_D \right\|_2 \cdot \left\| E_{x\sim S}\left[\Phi\left(x\right)\right] - E_{x\sim D}\left[\Phi\left(x\right)\right] \right\|_2\end{aligned}$$

where we use the Cauchy-Schwarz inequality and that $\left\| \hat{w} - w_D \right\|_2^2 = \left\| w_D - \hat{w} \right\|_2^2$. Divide both sides by using $\lambda\left(w_D - \hat{w}\right)$

$$\left\| \hat{w} - w_D \right\|_2 \leq \frac{1}{\lambda} \left\| E_{x\sim S}\left[\Phi\left(x\right)\right] - E_{x\sim D}\left[\Phi\left(x\right)\right] \right\|_2$$

3

To prove: $L_D(\hat{w}) - L_D(w_D) \leq (\hat{w} - w_D) \cdot (E_{x\sim S}\left[\Phi\left(x\right)\right] - E_{x\sim D}\left[\Phi\left(x\right)\right]) + \frac{\lambda}{2} \left\| w_D \right\|_2^2 - \frac{\lambda}{2} \left\| \hat{w} \right\|_2^2$

Proof:

We know that:

$$J_D(w) = \frac{\lambda}{2} \left\| w \right\|_2^2 + L_D(w) = U_D\left(w\right) + L_D(w)$$

where

$$U_D(w) = \frac{\lambda}{2} \left\| w \right\|_2^2$$

Also $\hat{w}$ minimizes J_D. $\therefore U_D\left(\hat{w}\right) + L_D(\hat{w}) \leq U_D\left(w\right) + L_D(w)$

$$\begin{aligned}L_D(\hat{w}) \leq& L_D(w) + U_D\left(w\right) - U_D\left(\hat{w}\right) = L_D(w) + U\left(-w\right) - U\left(-\hat{w}\right) + (w - \hat{w}) \cdot E_{x\sim D}\left[\Phi(x)\right] \\ \leq& L_D(w) + U_D\left(w\right) - U_D\left(\hat{w}\right) + (w - \hat{w}) \cdot (E_{x\sim D}\left[\Phi(x)\right] - E_{x\sim S}\left[\Phi(x)\right])\end{aligned}$$

The second equation can be obtained by shifting.

Using, $w = w_D$ and $U_D(w) = \frac{\lambda}{2} \left\| w \right\|_2^2$,

$$L_D(\hat{w}) \leq L_D(w_D) + \frac{\lambda}{2} \left\| w_D \right\|_2^2 - \frac{\lambda}{2} \left\| \hat{w} \right\|_2^2 + (w_D - \hat{w}) \cdot (E_{x\sim D}\left[\Phi(x)\right] - E_{x\sim S}\left[\Phi(x)\right])$$

$$L_D(\hat{w}) - L_D(w_D) \leq (\hat{w} - w_D) \cdot (E_{x\sim S}\left[\Phi(x)\right] - E_{x\sim D}\left[\Phi(x)\right]) + \frac{\lambda}{2} \left\| w_D \right\|_2^2 - \frac{\lambda}{2} \left\| \hat{w} \right\|_2^2$$

4

To Prove $L_D(\hat{w}) \leq \frac{1}{\lambda} \left\| E_{x\sim S}\left[\Phi\left(x\right)\right] - E_{x\sim D}\left[\Phi\left(x\right)\right]\right\|_2^2 + L_D(w) + \frac{\lambda}{2}\left\|w\right\|_2^2$

Proof We know that,

$$L_D(\hat{w}) - L_D(w_D) \leq (\hat{w} - w_D) \cdot (E_{x\sim S}\left[\Phi(x)\right] - E_{x\sim D}\left[\Phi(x)\right]) + \frac{\lambda}{2}\left\|w_D\right\|_2^2 - \frac{\lambda}{2}\left\|\hat{w}\right\|_2^2$$

and

$$\left\|\hat{w} - w_D\right\|_2 \leq \frac{1}{\lambda}\left\|E_{x\sim S}\left[\Phi\left(x\right)\right] - E_{x\sim D}\left[\Phi\left(x\right)\right]\right\|_2$$

By using Cauchy-Shwartz inequality,

$$(\hat{w} - w_D) \cdot (E_{x\sim S}\left[\Phi(x)\right] - E_{x\sim D}\left[\Phi(x)\right]) \leq \left\|\hat{w} - w_D\right\|_2 \cdot \left\|E_{x\sim S}\left[\Phi(x)\right] - E_{x\sim D}\left[\Phi(x)\right]\right\|_2$$

Substituting:

$$\begin{aligned} L_D(\hat{w}) - L_D(w_D) &\leq \left\|\hat{w} - w_D\right\|_2 \cdot \left\|E_{x\sim S}\left[\Phi(x)\right] - E_{x\sim D}\left[\Phi(x)\right]\right\|_2 + \frac{\lambda}{2}\left\|w_D\right\|_2^2 - \frac{\lambda}{2}\left\|\hat{w}\right\|_2^2 \\ &\leq \frac{1}{\lambda}\left\|E_{x\sim S}\left[\Phi(x)\right] - E_{x\sim D}\left[\Phi(x)\right]\right\|_2^2 + \frac{\lambda}{2}\left\|w_D\right\|_2^2 - \frac{\lambda}{2}\left\|\hat{w}\right\|_2^2 \\ &\leq \frac{1}{\lambda}\left\|E_{x\sim S}\left[\Phi(x)\right] - E_{x\sim D}\left[\Phi(x)\right]\right\|_2^2 + \frac{\lambda}{2}\left\|w_D\right\|_2^2 \end{aligned}$$

Since w_D minimizes $J_D(w)$, we know $\left[J_D\right]_{w=w_D} \leq \left[J_D\right]$arbitraryw

$\therefore \frac{\lambda}{2}\left\|w_D\right\|_2^2 + L_D(w_D) \leq \frac{\lambda}{2}\left\|w\right\|_2^2 + L_D(w)$

Using the above equations

$$\begin{aligned} L_D(\hat{w}) - L_D(w) &\leq \frac{1}{\lambda}\left\|E_{x\sim S}\left[\Phi(x)\right] - E_{x\sim D}\left[\Phi(x)\right]\right\|_2^2 + \frac{\lambda}{2}\left\|w\right\|_2^2 \\ L_D(\hat{w}) &\leq \frac{1}{\lambda}\left\|E_{x\sim S}\left[\Phi(x)\right] - E_{x\sim D}\left[\Phi(x)\right]\right\|_2^2 + L_D(w) + \frac{\lambda}{2}\left\|w\right\|_2^2 \end{aligned}$$

This proves the given equation.

5

Given: $\delta > 0$

To Prove: $\forall \delta > 0; \mathcal{L}_D(\hat{w}) \leq \inf_{w\in\mathbb{R}^N} \mathcal{L}_D(w) + \frac{\lambda}{2}\left\|w\right\|_2^2 + \frac{2r^2}{\lambda m}\left(1 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right)^2$ is valid with probability $1 - \delta$

Proof:

We have already proved that $\forall \delta$ with probability $1 - \delta$

$$\|E_{x\sim D}\left[\Phi(x)\right] - E_{x\sim S}\left[\Phi(x)\right]\|_2 \leq \sqrt{\frac{2r^2}{m}}\left(1 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right)$$

Also

$$\mathcal{L}_D(\hat{w}) \leq \frac{1}{\lambda}\|E_{x\sim S}\left[\Phi(x)\right] - E_{x\sim D}\left[\Phi(x)\right]\|_2^2 + \mathcal{L}_D(w) + \frac{\lambda}{2}\|w\|_2^2$$

Using the above two equations

$$\begin{aligned}
L_D(\hat{w}) \leq& \frac{1}{\lambda}\left\|E_{x\sim S}\left[\Phi(x)\right] - E_{x\sim D}\left[\Phi(x)\right]\right\|_2^2 + L_D(w) + \frac{\lambda}{2}\left\|w\right\|_2^2 \\
\leq& \frac{1}{\lambda}\left(\sqrt{\frac{2r^2}{m}}\left(1 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right)\right)^2 + L_D(w) + \frac{\lambda}{2}\left\|w\right\|_2^2 \\
=& L_D(w) + \frac{\lambda}{2}\left\|w\right\|_2^2 + \frac{2r^2}{\lambda m}\left(1 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right)^2
\end{aligned}$$

Since $\forall w \in \mathbb{R}^N$, $\inf_{w\in\mathbb{R}^N} \mathcal{L}_D(w) \leq \mathcal{L}_D(w)$ we can obtain a tight upper bound by using this.

$$L_D(\hat{w}) \leq \inf_{w\in\mathbb{R}^N} L_D(w) + \frac{\lambda}{2}\left\|w\right\|_2^2 + \frac{2r^2}{\lambda m}\left(1 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right)^2$$

Hence Proved.