Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Assignment Number: 2
Student Name: Gurpreet Singh
Roll Number: 150259
Date: March 18, 2018

QUESTION 1

From the discussion in class, we know about the following inequality, which I have used to prove some identities in this question.

If $a, b \geq 0$ and $a + b \geq c$, then for all $\eta \in [0, 1]$,

$$\eta \cdot a + (1 - \eta) \cdot b \quad \geq \quad c \cdot \min(\eta, 1 - \eta) \tag{1}$$

PART 1

$$\begin{aligned} L^{0-1}(\eta) &= \min_{\hat{y}} \underset{Y \sim \eta}{\mathbb{E}} \left[l^{0-1}(\hat{y}, Y) \right] \\ &= \min_{\hat{y}} \underset{Y \sim \eta}{\mathbb{E}} \left[\mathbb{I}[\hat{y} \neq Y] \right] \\ &= \min_{\hat{y}} \underset{Y \sim \eta}{\mathbb{E}} \left[\mathbb{I}[\hat{y} = 1] \mathbb{I}[Y = -1] + \mathbb{I}[\hat{y} = -1] \mathbb{I}[Y = 1] \right] \\ &= \min_{\hat{y}} (1 - \eta) \, \mathbb{I}[\hat{y} = 1] + \eta \, \mathbb{I}[\hat{y} = -1] \\ &\geq \min(\eta, 1 - \eta) \end{aligned}$$

The last inequality is from Equation 1. However, for an algorithm that predicts $\hat{y}$ as follows,

$$\hat{y} \quad = \quad \begin{cases} +1 & \text{if } \eta > 0.5 \\ -1 & \text{else} \end{cases}$$

the equality is satisfied. Therefore, we have a predictor that gives an error $\min(\eta, 1 - \eta)$, *i.e.* $L^{0-1}(\eta) = \min(\eta, 1 - \eta)$.

Similarly for $L^{\sigma}(\eta)$,

$$\begin{aligned} L^{\sigma}(\eta) &= \min_{\hat{y}} \underset{Y \sim \eta}{\mathbb{E}} \left[l^{\sigma}(\hat{y}, Y) \right] \\ &= \min_{\hat{y}} \underset{Y \sim \eta}{\mathbb{E}} \left[\frac{1}{1 + \exp(Y\hat{y})} \right] \\ &= \min_{\hat{y}} (1 - \eta) \frac{\exp(\hat{y})}{1 + \exp(\hat{y})} + \eta \frac{1}{1 + \exp(\hat{y})} \qquad \geq \quad \min(\eta, 1 - \eta) \end{aligned}$$

Again, the last inequality is from Equation 1. We can find a prediction function such that this is becomes an equality. The said prediction would be as follows

$$\hat{y} \quad = \quad \begin{cases} +\infty & \text{if } \eta > 0.5 \\ -\infty & \text{else} \end{cases}$$

Since for this predictor, the said inequality is indeed an equality, we can say that the minimum predictor gives $L^{\sigma}(\eta) = \min(\eta, 1 - \eta)$.

Hence, we have the result

$$L^{0-1}(\eta) = L^{\sigma}(\eta) = \min(\eta, 1 - \eta) \tag{2}$$

PART 2

From the previous part, we have

$$\begin{aligned} L^{0-1}(\hat{y},\eta) &= (1-\eta)\,\mathbb{I}\,[\,\hat{y}=1\,] + \eta\,\mathbb{I}\,[\,\hat{y}=-1\,] \\ L^{0-1}(\eta) &= \min{(\eta, 1-\eta)} \end{aligned}$$

Therefore, there are four cases for $L^{0-1}(\hat{y},\eta) - L^{0-1}(\eta)$,

$$L^{0-1}(\hat{y},\eta) - L^{0-1}(\eta) = \begin{cases} 1-2\eta & \text{if } \eta \le 0.5, \hat{y}=+1 \\ 0 & \text{if } \eta \le 0.5, \hat{y}=-1 \\ 0 & \text{if } \eta > 0.5, \hat{y}=+1 \\ 2\eta-1 & \text{if } \eta > 0.5, \hat{y}=-1 \end{cases}$$

From this, we can directly represent this in the form given in the question, *i.e.*

$$L^{0-1}(\hat{y},\eta) - L^{0-1}(\eta) = |\,2\eta-1\,| \cdot \mathbb{I}\,[\,\hat{y}\,(2\eta-1) < 0\,]$$

PART 3

From Part 1, we know

$$\begin{aligned} L^{0-1}(\text{sign}(\hat{y}),\eta) &= (1-\eta)\,\mathbb{I}\,[\,\text{sign}(\hat{y})=+1\,] + \eta\,\mathbb{I}\,[\,\text{sign}(\hat{y})=-1\,] \\ &= (1-\eta)\,\mathbb{I}\,[\,\hat{y}\ge 0\,] + \eta\,\mathbb{I}\,[\,\hat{y}<0\,] \end{aligned}$$

Now, consider the term $2L^{\sigma}(\hat{y},\eta) - L^{0-1}(\text{sign}(\hat{y}),\eta)$.

$$2L^{\sigma}(\hat{y},\eta) - L^{0-1}(\text{sign}(\hat{y}),\eta) = (1-\eta)\left[\frac{2\exp{(\hat{y})}}{1+\exp{(\hat{y})}} - \mathbb{I}\,[\,\hat{y}\ge 0\,]\right] + \eta\left[\frac{2}{1+\exp{(\hat{y})}} - \mathbb{I}\,[\,\hat{y}<0\,]\right]$$

Since for $x \ge 0$, $\frac{2\exp(x)}{1+\exp(x)} \ge 1$, we have $\frac{2\exp(\hat{y})}{1+\exp(\hat{y})} - \mathbb{I}\,[\,\hat{y}\ge 0\,] \ge 0$. Similary, if $x<0$, $\frac{2}{1+\exp(x)} > 1$. Therefore, we also have $\frac{2}{1+\exp(\hat{y})} - \mathbb{I}\,[\,\hat{y}<0\,] \ge 0$.

Hence, from Equation 1, we can say

$$\begin{aligned} 2L^{\sigma}(\hat{y},\eta) - L^{0-1}(\text{sign}(\hat{y}),\eta) &\ge \left[\frac{2\exp{(\hat{y})}+2}{1+\exp{(\hat{y})}} - \mathbb{I}\,[\,\hat{y}\ge 0\,] - \mathbb{I}\,[\,\hat{y}<0\,]\right]\cdot \min{(\eta,1-\eta)} \\ &= \min{(\eta,1-\eta)} \\ &= 2L^{\sigma}(\eta) - L^{0-1}(\eta) \end{aligned}$$

The last equality comes from Equation 2. Therefore, we have the desired result, that is

$$L^{0-1}(\text{sign}(\hat{y}),\eta) - L^{0-1}(\eta) \le 2\,(L^{\sigma}(\hat{y},\eta) - L^{\sigma}(\eta)) \tag{3}$$

PART 4

From part 3, equation 3, we have the following result

$$L^{0-1}(\text{sign}(\hat{y}),\eta) - L^{0-1}(\eta) \le 2\,(L^{\sigma}(\hat{y},\eta) - L^{\sigma}(\eta))$$

Suppose for some $f:\mathcal{X}\to\mathbb{R}$, we have $\hat{y}=f(\mathbf{x})$. Therefore, we can write

$$L^{0-1}(\text{sign}(f(\mathbf{x})),\eta) - L^{0-1}(\eta) \le 2\,(L^{\sigma}(f(\mathbf{x}),\eta) - L^{\sigma}(\eta))$$

Since this inequality exists for all $\mathbf{x} \in \mathcal{X}$, we can take expectation on both sides over $\mathbf{X} \sim \mathcal{D}$ without disturbing the inequality. Therefore

$$\mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{D}}\left[L^{0-1}(\text{sign}(f(\mathbf{X})),\eta) - L^{0-1}(\eta)\right] \quad \leq \quad \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{D}}\left[2\left(L^{\sigma}(f(\mathbf{X}),\eta) - L^{\sigma}(\eta)\right)\right]$$

Since both LHS and RHS are regret terms, we can replace them as follows,

$$\implies \mathcal{R}^{0-1}_{\mathcal{D}}\left[\text{sign}\circ f\right] \quad \leq \quad 2\mathcal{R}^{\sigma}_{\mathcal{D}}\left[f\right] \tag{4}$$

Hence, we have shown the inequality required.

PART 5

We have

$$l^{\sigma}_{\alpha} \quad = \quad \frac{1}{1+\exp\left(\alpha\cdot\hat{y}y\right)} \qquad \text{and} \qquad l^{\sigma} \quad = \quad \frac{1}{1+\exp\left(\hat{y}y\right)}$$

Therefore, we can write $l^{\sigma}_{\alpha}(\hat{y},y)$ as

$$l^{\sigma}_{\alpha}(\hat{y},y) \quad = \quad l^{\sigma}(\alpha\cdot\hat{y},y)$$

Therefore, we have

$$L^{\sigma}_{\alpha}(\hat{y},\eta) \quad = \quad L^{\sigma}(\alpha\cdot\hat{y},\eta)$$

Now, borrowing the results from part 1, we can write

$$L^{\sigma}_{\alpha}(\hat{y},\eta) \quad = \quad L^{\sigma}(\alpha\cdot\hat{y},\eta) \quad \geq \quad \min\left\{\eta, 1-\eta\right\}$$

Note. We can write this as the RHS is independent of the value of $\hat{y}$ and therefore always holds.

We can use the same predictor as we used in part 1 (since $\alpha > 0$) to show that there is in fact an optimal predictor which gives the lowest possible error ($\min\{\eta, 1-\eta\}$). The optimal predictor being

$$\hat{y} \quad = \quad \begin{cases} +\infty & \text{if } \eta > 0.5 \\ -\infty & \text{else} \end{cases}$$

Hence, we have

$$L^{\sigma}_{\alpha}(\eta) \quad = \quad \min\left\{\eta, 1-\eta\right\} \quad = \quad L^{\sigma}(\eta)$$

Now, from equation 4, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{D}}\left[L^{0-1}(\text{sign}(f(\mathbf{X})),\eta) - L^{0-1}(\eta)\right] \quad &\leq \quad 2\mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{D}}\left[L^{\sigma}(f(\mathbf{X}),\eta) - L^{\sigma}(\eta)\right] \\ \implies \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{D}}\left[L^{0-1}(\text{sign}(f(\mathbf{X})),\eta) - L^{0-1}(\eta)\right] \quad &\leq \quad 2\mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{D}}\left[L^{\sigma}(f(\mathbf{X}),\eta) - L^{\sigma}_{\alpha}(\eta)\right] \end{aligned}$$

Suppose we have some $g : \mathcal{X} \to \infty$ such that $f(\mathbf{x}) = \alpha\cdot g(\mathbf{x})$. Clearly, such a g will exist for all f. Therefore, we can replace f by $\alpha\cdot g$ as follows,

$$\mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{D}}\left[L^{0-1}(\text{sign}(\alpha\cdot g(\mathbf{X})),\eta) - L^{0-1}(\eta)\right] \quad \leq \quad 2\mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{D}}\left[L^{\sigma}(\alpha\cdot g(\mathbf{X}),\eta) - L^{\sigma}_{\alpha}(\eta)\right]$$

Also, since $\alpha > 0$, we have $\text{sign}\alpha\cdot g(\mathbf{x}) = \text{sign}g(\mathbf{x})$, and $L^{\sigma}(\alpha\cdot g(\mathbf{x}),\eta) = L^{\sigma}_{\alpha}(g(\mathbf{x}),\eta)$ for all $\mathbf{x} \in \mathcal{X}$. Hence, we can write this as

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{D}}\left[L^{0-1}(\text{sign}(g(\mathbf{X})),\eta) - L^{0-1}(\eta)\right] \quad &\leq \quad 2\mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{D}}\left[L^{\sigma}_{\alpha}(g(\mathbf{X}),\eta) - L^{\sigma}_{\alpha}(\eta)\right] \\ \implies \mathcal{R}^{0-1}_{\mathcal{D}}\left[\text{sign}\circ g\right] \quad &\leq \quad 2\mathcal{R}^{\sigma;\alpha}_{\mathcal{D}}\left[g\right] \end{aligned}$$

Therefore, we have the exact same regret transfer bound for l^{σ}_{α} as for l^{σ} with respect to l^{0-1}.

Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Assignment Number: 2
Student Name: Gurpreet Singh
Roll Number: 150259
Date: March 18, 2018

QUESTION 2

The hypothesis class we are working with is $\mathcal{V} = \{v \geq 0\}$. Each $v \in \mathcal{V}$ splits the set of people (expenditures) into two classes.

We first define $g : \mathcal{I}^n \to \mathbb{R}$ as follows

$$g(\mathcal{S}) \quad = \quad \sup_{v \in \mathcal{V}} \frac{1}{n} \cdot |\,\{i \in \mathcal{S} : E(i) \leq v\}\,| - \frac{1}{N} \cdot |\,\{i \in \mathcal{I} : E(i) \leq v\}\,|$$

It is clear that the desired quantity $\left|\, \frac{1}{N} \cdot \left|\, \left\{ i \in \mathcal{I} : E(i) \leq \hat{v}_{\mathcal{S}}^{\kappa} \right\} \right| - \kappa \,\right|$ is upper bounded by $g(\mathcal{S})$. Therefore, it is sufficient to upper bound (with high probability) $g(\mathcal{S})$ in order to get the derired concentration bound.

Claim 2.1. $g(\mathcal{S})$ is c-stable

Proof. Define $\mathcal{S}_i = \mathcal{S} \backslash \{I_i\} \bigcup \{I'_i\}$. Then, we want to prove $g(\mathcal{S}) - g(\mathcal{S}_i)$. Suppose we define $g_v : \mathcal{I}^n \to \mathbb{R}$ as follows

$$g_v(\mathcal{S}) \quad = \quad \left|\, \frac{1}{n} \cdot |\,\{i \in \mathcal{S} : E(i) \leq v\}\,| - \frac{1}{N} \cdot |\,\{i \in \mathcal{I} : E(i) \leq v\}\,| \,\right|$$

Then, we can write

$$\begin{aligned} |\, g_v(\mathcal{S}) - g_v(\mathcal{S}_i) \,| \quad &= \quad \Bigg|\, \left|\, \frac{1}{n} \cdot |\,\{j \in \mathcal{S} : E(j) \leq v\}\,| - \frac{1}{N} \cdot |\,\{j \in \mathcal{I} : E(j) \leq v\}\,| \,\right| \; - \; \ldots \\ &\qquad \ldots \; - \; \left|\, \frac{1}{n} \cdot |\,\{j \in \mathcal{S}_i : E(j) \leq v\}\,| - \frac{1}{N} \cdot |\,\{j \in \mathcal{I} : E(j) \leq v\}\,| \,\right| \,\Bigg| \\ &\leq \quad \left|\, \frac{1}{n} \cdot |\,\{j \in \mathcal{S} : E(j) \leq v\}\,| - \frac{1}{n} \cdot |\,\{j \in \mathcal{S}_i : E(j) \leq v\}\,| \,\right| \\ &= \quad \left|\, \frac{1}{n} \Big(\mathbb{I}\,[\, E(I_i) \leq v \,] - \mathbb{I}\,[\, E(I'_i) \leq v \,] \Big) \,\right| \end{aligned}$$

$$\implies |\, g_v(\mathcal{S}) - g_v(\mathcal{S}_i) \,| \quad \leq \quad \frac{1}{n} \tag{5}$$

Now, without loss of generality, assume $g(\mathcal{S}) = g_{f_1}(\mathcal{S})$ and $g(\mathcal{S}_i) = g_{f_2}(\mathcal{S}_i)$. Hence, we can write

$$\begin{aligned} g(\mathcal{S}) - g(\mathcal{S}_i) \quad &= \quad g_{f_1}(\mathcal{S}) - g_{f_2}(\mathcal{S}_i) \\ &= \quad \underbrace{\big(g_{f_1}(\mathcal{S}) - g_{f_1}(\mathcal{S}_i) \big)}_{(1)} + \underbrace{\big(g_{f_1}(\mathcal{S}_i) - g_{f_2}(\mathcal{S}_i) \big)}_{(2)} \end{aligned}$$

The (1) term in the above expression is less than equal to $1/n$ (using Equation 5), and the (2) term is non-positive, since $g_f(\mathcal{S})$ is maximum at $f = f_2$. Also, the same inequalities will hold if we consider $g(\mathcal{S}_i) - g(\mathcal{S})$. Therefore, we can write

$$|\, g(\mathcal{S}) - g(\mathcal{S}_i) \,| \quad \leq \quad \frac{1}{n} \tag{6}$$

□

Since the function $g(\mathcal{S})$ is $1/n$-stable, we can use McDiarmid's inequality to write

$$\mathbb{P}\left[g(\mathcal{S}) > \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^n}[\, g(\mathcal{S})\,] + \epsilon \right] \quad \leq \quad \exp\left(-2n\epsilon^2\right) \tag{7}$$

We now wish to bound $\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^n}[\, g(\mathcal{S})\,]$. If we rewrite $g(\mathcal{S})$ as follows

$$g(\mathcal{S}) \quad = \quad \sup_{v\in\mathcal{V}} \left| \frac{1}{n}\sum_{i=1}^{N} f_v(i) - \mathop{\mathbb{E}}_{i}[\, f(i)\,] \right|$$

where

$$f_v(i) \quad = \quad \mathbb{I}[\, E(i) \leq v\,]$$

then we can make use of Symmetrization (as discussed in class), and therefore find an upper bound on $\mathbb{E}[\, g(\mathcal{S})\,]$ (using Rademacher Variables). Hence, we have

$$\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^n}[\, g(\mathcal{S})\,] \quad \leq \quad 2R_n(\mathcal{F}) \tag{8}$$

where $\mathcal{F} = \{f_v\}_{v\in\mathcal{V}}$ is our transformed function class. Although the size of this function class is infinite, we can still work with class as the VC dimension of this class is 2. This comes from the discussion in class on the simple binary classification problem with a random variable splitting the real line into two parts, which is exactly what the function class $\mathcal{F}$ is achieving, although the labelling, in this case is $\{0,1\}$ instead of $\{1,-1\}$.

Therefore, for a sample $(\mathcal{S})$ of size n, we can easily find the size of the $\mathcal{S}$-restriction of $\mathcal{F}$ to be equal to $n+1$. Also, the maximum norm is $c \leq \sqrt{n}$. Hence, we can apply Massarts Finite Class Lemma to get

$$\begin{aligned} R_n(\mathcal{F}) \quad &\leq \quad \frac{c\sqrt{2\log\left(|\,\mathcal{F}|_{\mathcal{S}}\,|\right)}}{n} \\ &= \quad \sqrt{\frac{2\log\left(n+1\right)}{n}} \end{aligned}$$

Therefore, using Equations 7 and 8, we have

$$\mathbb{P}\left[g(\mathcal{S}) > 2\sqrt{\frac{2\log\left(n+1\right)}{n}} + \epsilon \right] \quad \leq \quad \exp\left(-2n\epsilon^2\right)$$

Since $g_{\hat{v}^{\kappa}_{\mathcal{S}}}(\mathcal{S}) < g(\mathcal{S})$, we can say

$$\mathbb{P}\left[\left| \frac{1}{n}\cdot|\{i\in\mathcal{S} : E(i) \leq \hat{v}^{\kappa}_{\mathcal{S}}\}| - \frac{1}{N}\cdot|\{i\in\mathcal{I} : E(i) \leq \hat{v}^{\kappa}_{\mathcal{S}}\}| \right| > 2\sqrt{\frac{2\log\left(n+1\right)}{n}} + \epsilon \right] \quad \leq \quad \exp\left(-2n\epsilon^2\right)$$

Since the first term within the expression to bound is equal to κ (by the design of the Algorithm), we have the desired bound

$$\mathbb{P}\left[\left| \frac{1}{N}\cdot|\{i\in\mathcal{I} : E(i) \leq \hat{v}^{\kappa}_{\mathcal{S}}\}| - \kappa \right| > 2\sqrt{\frac{2\log\left(n+1\right)}{n}} + \epsilon \right] \quad \leq \quad \exp\left(-2n\epsilon^2\right)$$

or alternatively

$$\forall\,\delta\in(0,1)\,, \quad \mathbb{P}\left[\left| \frac{1}{N}\cdot|\{i\in\mathcal{I} : E(i) \leq \hat{v}^{\kappa}_{\mathcal{S}}\}| - \kappa \right| > 2\sqrt{\frac{2\log\left(n+1\right)}{n}} + \sqrt{\frac{\log\left(1/\delta\right)}{2n}} \right] \quad \leq \quad \delta$$

Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Assignment Number: 2
Student Name: Gurpreet Singh
Roll Number: 150259
Date: March 18, 2018

QUESTION 3

It can help to rewrite the FAIR(.) using the following observation

$$\begin{aligned}
\underset{X,Y\sim\mathcal{D}}{\mathbb{E}}\left[l(f(X),Y)\,\middle|\,X\in\mathcal{X}\right] &= \sum_{\mathcal{D}} l(f(X),Y)\cdot\mathbb{P}\left[X,Y\,\middle|\,X\in\mathcal{X}\right]\\
&= \sum_{\mathcal{D}} l(f(X),Y)\cdot\frac{\mathbb{P}\left[X\in\mathcal{X}\,\middle|\,X,Y\right]\mathbb{P}\left[X,Y\right]}{\sum_{\mathcal{D}}\mathbb{P}\left[X\in\mathcal{X}\,\middle|\,X,Y\right]\mathbb{P}\left[X,Y\right]}\\
&= \sum_{\mathcal{D}} l(f(X),Y)\cdot\frac{\mathbb{P}\left[X\in\mathcal{X}\,\middle|\,X,Y\right]\mathbb{P}\left[X,Y\right]}{\mathbb{P}\left[X\in\mathcal{X}\right]}
\end{aligned}$$

Note. The summation in the above expressions could be an integral if $\mathcal{D}$ is continuous, however it would only bring about notational changes, and no difference to the analysis.

The first equality is from the definition of expectation and the second equality is from rewriting the probability using Bayes Rule. Now given X, the probability of $X\in\mathcal{X}$ is simply equal to $\mathbb{I}\left[X\in\mathcal{X}\right]$. Hence, we can rewrite this as

$$\begin{aligned}
\underset{X,Y\sim\mathcal{D}}{\mathbb{E}}\left[l(f(X),Y)\,\middle|\,X\in\mathcal{X}\right] &= \sum_{\mathcal{D}} l(f(X),Y)\cdot\mathbb{I}\left[X\in\mathcal{X}\right]\cdot\frac{\mathbb{P}\left[X,Y\right]}{\mathbb{P}\left[X\in\mathcal{X}\right]}\\
&= \frac{1}{\mathbb{P}\left[X\in\mathcal{X}\right]}\cdot\underset{X,Y\sim\mathcal{D}}{\mathbb{E}}\left[l(f(X),Y)\cdot\mathbb{I}\left[X\in\mathcal{X}\right]\right]
\end{aligned}$$

Note. The denominator is an integral/summation over the complete distribution, and is therefore independent of the expectation. Hence, we can bring it out of the expectation.

Now substituting this in the FAIR(.) expression, we get

$$\begin{aligned}
\text{FAIR}(f) &= \left|\frac{1}{\mathbb{P}\left[X\in\mathcal{X}_1\right]}\cdot\underset{X,Y\sim\mathcal{D}}{\mathbb{E}}\left[l(f(X),Y)\cdot\mathbb{I}\left[X\in\mathcal{X}_1\right]\right] - \dots\right.\\
&\qquad\left.\dots - \frac{1}{\mathbb{P}\left[X\in\mathcal{X}_2\right]}\cdot\underset{X,Y\sim\mathcal{D}}{\mathbb{E}}\left[l(f(X),Y)\cdot\mathbb{I}\left[X\in\mathcal{X}_2\right]\right]\right|
\end{aligned}$$

Since $\mathbb{P}\left[X\in\mathcal{X}_1\right] = 1/2 = \mathbb{P}\left[X\in\mathcal{X}_2\right]$, we can rewrite this as

$$\text{FAIR}(f) = 2\left|\underset{X,Y\sim\mathcal{D}}{\mathbb{E}}\left[l(f(X),Y)\left(\mathbb{I}\left[X\in\mathcal{X}_1\right]-\mathbb{I}\left[X\in\mathcal{X}_2\right]\right)\right]\right| \tag{9}$$

Suppose we ignore the constant 2 and the absolute function in the above expression, and temporarily redefine the notion of fairness as below

$$\text{FAIR}(f) = \underset{X,Y\sim\mathcal{D}}{\mathbb{E}}\left[l(f(X),Y)\left(\mathbb{I}\left[X\in\mathcal{X}_1\right]-\mathbb{I}\left[X\in\mathcal{X}_2\right]\right)\right]$$

Now, we propose an estimator for this redifed notion of fairness, with respect to a function f, as follows

$$\overline{\text{FAIR}}(f,\mathcal{S}) = \frac{1}{N}\sum_{X,Y\in\mathcal{S}} l(f(X),Y)\left(\mathbb{I}\left[X\in\mathcal{X}_1\right]-\mathbb{I}\left[X\in\mathcal{X}_2\right]\right) \tag{10}$$

where $\mathcal{S}$ is sample of size N, such that $\mathcal{S} \sim \mathcal{D}^N$

However, we need to show that this indeed is a good estimator for the notion of fairness we have defined. To see this, suppose we define $g_f(.)$ as follows

$$g_f(\mathcal{S}) \quad = \quad \left| \overline{\text{FAIR}}(f, \mathcal{S}) - \text{FAIR}(f) \right|$$

and

$$g(\mathcal{S}) \quad = \quad \sup_{f \in \mathcal{F}} \left\{ g_f(\mathcal{S}) \right\}$$

Claim 2.2. $g(.)$ is a c-stable function

Proof. Suppose we have a sample $\mathcal{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \ldots (\mathbf{x}_N, y_N)\}$.

Define $\mathcal{S}_n = \mathcal{S} \setminus \{(\mathbf{x}_n, y_n)\} \bigcup \left\{ \left(\mathbf{x}'_n, y'_n\right) \right\}$, then we have

$$\begin{aligned} \left| g_f(\mathcal{S}) - g_f(\mathcal{S}_n) \right| \quad &= \quad \left| \left| \overline{\text{FAIR}}(f, \mathcal{S}) - \text{FAIR}(f) \right| - \left| \overline{\text{FAIR}}(f, \mathcal{S}_n) - \text{FAIR}(f) \right| \right| \\ &\leq \quad \left| \overline{\text{FAIR}}(f, \mathcal{S}) - \overline{\text{FAIR}}(f, \mathcal{S}_n) \right| \\ &= \quad \frac{1}{N} \left| l(f(\mathbf{x}_n), y_n) - l(f(\mathbf{x}'_n), y'_n) \right| \end{aligned}$$

Since $l : \mathbb{R} \times \mathcal{Y} \rightarrow [0, B]$, we have

$$g_f(\mathcal{S}) - g_f(\mathcal{S}_n) \quad \leq \quad \frac{B}{N} \tag{11}$$

However, we need to bound the difference between $g(\mathcal{S})$ and $g(\mathcal{S}_n)$. From the definition of $g(.)$, we have

$$g(\mathcal{S}) - g(\mathcal{S}_n) \quad = \quad \sup_{f \in \mathcal{F}} g_f(\mathcal{S}) - \sup_{f' \in \mathcal{F}} g_{f'}(\mathcal{S}_n)$$

Without loss of generality, we can say $g_f(\mathcal{S})$ is maximum for $f_1 \in \mathcal{F}$ and $g_f(\mathcal{S}_n)$ is maximum for $f_2 \in \mathcal{F}$ as well. Therefore, we can say

$$\begin{aligned} g(\mathcal{S}) - g(\mathcal{S}_n) \quad &= \quad g_{f_1}(\mathcal{S}) - g_{f_2}(\mathcal{S}_n) \\ &\leq \quad \big(g_{f_1}(\mathcal{S}) - g_{f_1}(\mathcal{S}_n)\big) + \big(g_{f_1}(\mathcal{S}_n) - g_{f_2}(\mathcal{S}_n)\big) \end{aligned}$$

The second term is non-positive from the definition of f_2, which maximizes $g_f(\mathcal{S}_n)$. The first term is smaller than B/N (using Equation 11). Hence, we have

$$g(\mathcal{S}) - g(\mathcal{S}_n) \quad \leq \quad \frac{B}{N}$$

Using symmetrization and the fact that $\mathcal{S}$ and $\mathcal{S}_n$ are similar, we have the above inequality for $g(\mathcal{S}_n) - g(\mathcal{S})$. Therefore, we have

$$\left| g(\mathcal{S}) - g(\mathcal{S}_n) \right| \quad \leq \quad \frac{B}{N} \tag{12}$$

□

Since we have the fact that $g(\mathcal{S})$ is B/N-stable, we can use McDiarmid's Inequality to say

$$\text{if } N > -\frac{B^2}{2\epsilon^2} \log(\delta), \qquad \mathbb{P}\left[g(\mathcal{S}) - \mathop{\mathbb{E}}_{\mathcal{S}} [\, g(\mathcal{S}) \,] > \epsilon \right] \quad \leq \quad \delta \tag{13}$$

where $N = |\, \mathcal{S} \,|$

Note. We do not have $\delta/2$ in the expression for the lower bound on the sample size simply because we are only considering right tail, and not the let tail at this time.

From the definition of $g(.)$, we have for a given training sample $\mathcal{S} \sim \mathcal{D}^N$, for all functions $f \in \mathcal{F}$,

$$g_f(\mathcal{S}) \quad = \quad \left| \overline{\text{FAIR}}(f, \mathcal{S}) - \text{FAIR}(f) \right| \quad \leq \quad g(\mathcal{S})$$

Therefore, this also holds for $\hat{f}_{\mathcal{S}'}$ for some sample $S' \sim \mathcal{D}^M$, where $M \in \mathbb{N}$. This suggests that the inequality we derived in Equation 12 can be used to solve our original problem, *i.e.* estimating $\text{FAIR}(f_{\mathcal{S}})$.

However, we also need to bound the value of $\mathbb{E}\left[\, g(\mathcal{S}) \,\right]$, in order to bound the estimate of fairness we have derived.

$$\underset{\mathcal{S}\sim\mathcal{D}^N}{\mathbb{E}} \left[\, g(\mathcal{S}) \,\right] \quad = \quad \underset{\mathcal{S}\sim\mathcal{D}^N}{\mathbb{E}} \left[\sup_{f\in\mathcal{F}} \left\{ \overline{\text{FAIR}}(f, \mathcal{S}) - \text{FAIR}(f) \right\} \right]$$

From the discussion in class (Symmetrization and Rademacher Complexity), we know that the above expectation will be lesser than 2 times the Rademacher Complexity of the Function Class transformed by the loss function. More formally,

$$\mathbb{E}\left[\, g(\mathcal{S}) \,\right] \quad \leq \quad 2\, R_N \left(l^d \circ \mathcal{F} \right) \tag{14}$$

where

$$l^d(X, f(X), Y) \quad = \quad l(f(X), Y) \left(\mathbb{I}\left[\, X \in \mathcal{X}_1 \,\right] - \mathbb{I}\left[\, X \in \mathcal{X}_2 \,\right] \right)$$

Therefore, it suffices to bound the Rademacher complexity $\left(R_N \left(l^d \circ \mathcal{F} \right) \right)$ for our problem.

Claim 2.3. $R_N(l^d \circ \mathcal{F}) = R_N(l \circ \mathcal{F})$

Proof.

$$\begin{aligned} R_N(l^d \circ \mathcal{F}) \quad &= \quad \underset{\epsilon_n, \mathbf{x}_n, y_n}{\mathbb{E}} \left[\sup_{f\in\mathcal{F}} \sum_{n=1}^{N} \epsilon_i \cdot l^d(\mathbf{x}_n, f(\mathbf{x}_n), y_n) \right] \\ &= \quad \underset{\epsilon_n, \mathbf{x}_n, y_n}{\mathbb{E}} \left[\sup_{f\in\mathcal{F}} \sum_{n=1}^{N} \epsilon_i \cdot l(f(\mathbf{x}_n), y_n) \left(\mathbb{I}\left[\, \mathbf{x}_n \in \mathcal{X}_1 \,\right] - \mathbb{I}\left[\, \mathbf{x}_n \in \mathcal{X}_2 \,\right] \right) \right] \end{aligned}$$

Suppose we separate the samples $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$ into two sets $\mathcal{S}^+$ and $\mathcal{S}^-$ such that for all $i \in \mathcal{S}^+$, $\mathbb{I}\left[\, \mathbf{x}_i \in \mathcal{X}_1 \,\right] = 1$, and for all $j \in \mathcal{S}^-$, $\mathbb{I}\left[\, \mathbf{x}_j \in \mathcal{X}_2 \,\right] = 1$. Then, we can write

$$R_N(l^d \circ \mathcal{F}) \quad = \quad \underset{\epsilon_n, \mathbf{x}_n, y_n}{\mathbb{E}} \left[\sup_{f\in\mathcal{F}} \left\{ \sum_{i\in\mathcal{S}^+} \epsilon_i \cdot l(f(\mathbf{x}_i), y_i) + \sum_{j\in\mathcal{S}^-} (-\epsilon_j) \cdot l(f(\mathbf{x}_j), y_j) \right\} \right]$$

Since $-\epsilon_j$ is also a Rademacher variable (discussed in class), we can alternatively write this as

$$R_N(l^d \circ \mathcal{F}) \quad = \quad \underset{\epsilon'_n, \mathbf{x}_n, y_n}{\mathbb{E}} \left[\sup_{f\in\mathcal{F}} \sum_{n=1}^{N} \epsilon'_n \cdot l(f(\mathbf{x}_n), y_n) \right]$$

The RHS is exactly the expression for $R_N(l \circ \mathcal{F})$. Hence

$$R_N(l^d \circ \mathcal{F}) = R_N(l \circ \mathcal{F}) \tag{15}$$

□

Also, since we know that the loss function $l(n)$ is L-Lipschitz, using Ledoux-Talagrand Contraction Inequality, we can write

$$R_N(l \circ \mathcal{F}) \quad \leq \quad L \cdot R_N(\mathcal{F}) \tag{16}$$

Now, using Equations 13, 14, 15 and 16, we can write

$$\mathbb{E}\left[\, g(\mathcal{S}) \,\right] \quad \leq \quad 2 \cdot R_N \left(l^d \circ \mathcal{F} \right) \quad = \quad 2 \cdot R_N \left(\mathcal{F} \right) \quad \leq \quad 2L \cdot R_N \left(\mathcal{F} \right)$$

Therefore, we can say

$$\text{if } N > -\frac{B^2}{2\epsilon^2} \log\left(\delta\right), \qquad \mathbb{P}\left[\, g(\mathcal{S}) > 2L \cdot R_N\left(\mathcal{F}\right) + \epsilon \,\right] \quad \leq \quad \delta$$

Also, from the definition of $g(.)$, we have for any function $f \in \mathcal{F}$, $g_f(\mathcal{S}) \leq g(\mathcal{S})$. Therefore, we can further substitute the above equation to say

$$\text{if } N > -\frac{B^2}{2\epsilon^2} \log\left(\delta\right), \qquad \forall\, f \in \mathcal{F}, \quad \mathbb{P}\left[\, g_f(\mathcal{S}) > 2L \cdot R_N\left(\mathcal{F}\right) + \epsilon \,\right] \quad \leq \quad \delta \tag{17}$$

Since the above bound holds for all $f \in \mathcal{F}$, we can put $f = \hat{f}_{\mathcal{S}'}$ for some training sample $\mathcal{S}' \sim \mathcal{D}^M$. Substituting the expression of $g_f(\mathcal{S})$, we finally have

$$\text{if } N > -\frac{B^2}{2\epsilon^2} \log\left(\delta\right), \qquad \mathbb{P}\left[\left| \overline{\text{FAIR}}\left(\hat{f}_{\mathcal{S}'}, \mathcal{S} \right) - \text{FAIR}\left(\hat{f}_{\mathcal{S}'} \right) \right| > 2L \cdot R_N\left(\mathcal{F}\right) + \epsilon \right] \quad \leq \quad \delta$$

Since $|\, x - y \,| > |\, |\, x \,| - |\, y \,| \,|$, we further have

$$\text{if } N > -\frac{B^2}{2\epsilon^2} \log\left(\delta\right), \quad \mathbb{P}\left[\left| \left| \overline{\text{FAIR}}\left(\hat{f}_{\mathcal{S}'}, \mathcal{S} \right) \right| - \left| \text{FAIR}\left(\hat{f}_{\mathcal{S}'} \right) \right| \right| > 2L \cdot R_N\left(\mathcal{F}\right) + \epsilon \right] \quad \leq \quad \delta \tag{18}$$

The above bound suggests that $\overline{\text{FAIR}}(f, \mathcal{S})$ is, in fact, a good estimate for the desired quantity $\text{FAIR}(f)$. Also, it is possible to estimate $\text{FAIR}\left(\hat{f}_{\mathcal{S}'} \right)$ using $\mathcal{S}'$ itself, simply by setting $\mathcal{S} = \mathcal{S}'$.

Now, we can bring back the constant 2 and the absolute function we claimed to ignore earlier. However, this will not change the analysis as there is already and abolute over both the estimator and the actual assumed measure of fairness in Equation 18. Therefore, to conclude, the proposed estimator for $\text{FAIR}\left(\hat{f}_{\mathcal{S}} \right)$ is given as

$$\overline{\text{FAIR}}\left(\hat{f}_{\mathcal{S}}, \mathcal{S} \right) \quad = \quad \left| \frac{2}{N} \sum_{X, Y \in \mathcal{S}} l(\hat{f}_{\mathcal{S}}(X), Y) \left(\mathbb{I}\left[\, X \in \mathcal{X}_1 \,\right] - \mathbb{I}\left[\, X \in \mathcal{X}_2 \,\right] \right) \right|$$

The above estimator enjoys the concentration bound given in Equation 18, as shown earlier. Also, if the function class $\mathcal{F}$ has a small Rademacher Complexity (condition satisfied in the question), then we can claim that the estimator is indeed a good approximation of the measure of fairness, as defined in the question. We can finally say, for small Rademacher Complexity of $\mathcal{F}$, we have

$$\mathbb{P}\left[\left| \overline{\text{FAIR}}\left(\hat{f}_{\mathcal{S}'}, \mathcal{S} \right) - \text{FAIR}\left(\hat{f}_{\mathcal{S}'} \right) \right| > 4L \cdot R_N\left(\mathcal{F}\right) + \epsilon \right] \quad \leq \quad \exp\left(-\frac{N\epsilon^2}{2B^2} \right)$$

or alternatively

$$\forall\, \delta \in (0, 1), \quad \mathbb{P}\left[\left| \overline{\text{FAIR}}\left(\hat{f}_{\mathcal{S}'}, \mathcal{S} \right) - \text{FAIR}\left(\hat{f}_{\mathcal{S}'} \right) \right| > 4L \cdot R_N\left(\mathcal{F}\right) + B\sqrt{\frac{2\log\left(1/\delta\right)}{N}} \right] \quad \leq \quad \delta$$