Rademacher Average and Covering Number

Ruixin Guo

Department of Computer Science
Kent State University

April 17, 2023

Contents

Recall

We want to find a bound for $R^{\text{true}}(f) - R^{\text{emp}}(f)$ where f is from a function class $\mathcal{F}$.

If $|\mathcal{F}| = N$ is finite, the bound can be written as

$$P\left[\sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) \leq \sqrt{\frac{\log N + \log\frac{1}{\delta}}{2m}}\right] \geq 1-\delta$$

If $|\mathcal{F}|$ is infinite, we need to find a finite measure for the capacity of $\mathcal{F}$. Growth Function and VC Dimension are two classes of such measure and they are closely related.

Growth Function: The maximum number of ways into which m points can be classified by $\mathcal{F}$, denoted as $S_{\mathcal{F}}(m)$.

Shattering: We say $\mathcal{F}$ shatters an m-point dataset if $S_{\mathcal{F}} = 2^m$. That is, for an arbitrary way of classifying m points, there always exists one f in $\mathcal{F}$ that can generate it.

VC Dimension: The VC Dimension of a function class $\mathcal{F}$ is the largest h such that $S_{\mathcal{F}} = 2^h$, i.e., the maximum number of points that $\mathcal{F}$ can shatter.

Recall

VC Theorem: The bound for $R^{\text{true}}(f) - R^{\text{emp}}(f)$ in terms of Growth Function is,

$$P\left[\sup_{f\in\mathcal{F}}\left(R^{\text{true}}(f) - R^{\text{emp}}(f)\right) \leq 2\sqrt{2\frac{\log S_{\mathcal{F}}(2m) + \log\frac{4}{\delta}}{m}}\right] \geq 1-\delta$$

Sauer's Lemma shows an upper bound of the growth function. Supposing the VC Dimension of $\mathcal{F}$ is h, and there are m points to be classified, then the growth function can be written as

$$S_{\mathcal{F}}(m) \leq \sum_{i=0}^{h}\binom{m}{i}$$

In particular, when $m \geq h$, we further have

$$S_{\mathcal{F}}(m) \leq (\frac{em}{h})^h$$

By combining Sauer's Lemma and VC Theorem we get

$$P\left[\sup_{f\in\mathcal{F}}\left(R^{\text{true}}(f) - R^{\text{emp}}(f)\right) \leq 2\sqrt{2\frac{h\log\frac{2em}{h} + \log\frac{4}{\delta}}{m}}\right] \geq 1-\delta$$

Other Types of Capacity Measure

The capacity measure is a metric for the learning ability (complexity/expressiveness/richness...) of a function class $\mathcal{F}$.

Some types of Capacity Measure:

- Growth Function and VC dimension
- VC entropy
- Covering Numbers
- Rademacher Average

Growth Function and VC dimension are independent of data distribution. The bound based on them may be loose for most distributions.

The other 3 measures are data dependent, which means they can generate tighter bounds. We will focus on Rademacher Average and Covering Numbers.

Contents

Rademacher Average

One way to measure complexity is to see how functions from the class $\mathcal{F}$ can classify random noise.

Suppose the dataset $S = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ has m samples where $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$ (Note that y_i is not in $\{0, 1\}$ anymore). Let each $f \in \mathcal{F}$ be a binary classification function such that $f(x_i) \in \{-1, 1\}$ for $i = 1, 2, ..., m$.

We can write the empirical risk function as

$$
\begin{aligned}
R^{\text{emp}}(f) &= \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{f(x_i) \neq y_i} \\
&= \frac{1}{m} \sum_{i=1}^{m} \begin{cases} 1 & (f(x_i), y_i) = (1, -1) \text{ or } (-1, 1) \\ 0 & (f(x_i), y_i) = (1, 1) \text{ or } (-1, -1) \end{cases} \\
&= \frac{1}{m} \sum_{i=1}^{m} \frac{1 - y_i f(x_i)}{2} \\
&= \frac{1}{2} - \frac{1}{2m} \sum_{i=1}^{m} y_i f(x_i)
\end{aligned}
$$

The term $\frac{1}{m} \sum_{i=1}^{m} y_i f(x_i)$ can be interpreted as the correlation of the predictions $f(x_i)$ with the labels y_i. Since $\frac{1}{m} \sum_{i=1}^{m} y_i f(x_i) = 1 - 2R^{\text{emp}}(f)$, the greater $\frac{1}{m} \sum_{i=1}^{m} y_i f(x_i)$ makes the smaller $R^{\text{emp}}(f)$, which means $\mathcal{F}$ has stronger ability to learn.

Rademacher Average

Thus our goal is to find an $f \in \mathcal{F}$ satisfying

$$\underset{f \in \mathcal{F}}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^{m} y_i f(x_i)$$

Now assume the labels satisfy a specific distribution, i.e., each label is a Rademacher random variable.

Definition (Rademacher Variable): A random variable σ is called Rademacher random variable if $P(\sigma = 1) = P(\sigma = -1) = 1/2$.

Let $\sigma_1, \sigma_2, ..., \sigma_m$ be iid Rademacher random variables. Replacing $y_1, y_2, ..., y_m$ by $\sigma_1, \sigma_2, ..., \sigma_m$, we get

$$\underset{f \in \mathcal{F}}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i)$$

Instead of select the $f \in \mathcal{F}$ that correlates best with labels, this selects the $f \in \mathcal{F}$ that correlates best with random noise variables σ_i. To measure how well F correlate with random noise, we take the expectation of this correlation over the random variables σ_i:

Definition (Empirical Rademacher Average): For a class $\mathcal{F}$ of functions, on a given dataset S, the Empirical Rademacher Average is defined as

$$\hat{\mathcal{R}}_S(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i)\right)$$

Rademacher Average

By our definition of Empirical Rademacher Average $\hat{\mathcal{R}}_S(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} (\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i))$, we have $0 \leq \hat{\mathcal{R}}_S(\mathcal{F}) \leq 1$. Consider two extreme cases:

- When $S_m(\mathcal{F}) = 1$, we only have one group of f and can only generate one way of classification. In this case $\hat{\mathcal{R}}_S(\mathcal{F}) = 0$ since the max term disappears.
- When $S_m(\mathcal{F}) = 2^m$, $\mathcal{F}$ shatters S. In this case $\hat{\mathcal{R}}_S(\mathcal{F}) = 1$.

Remember that the dataset S is sampled from an unknown distribution D. The expectation of the empirical Rademacher average of all $S \sim D$ is called the Rademacher Average:

Definition (Rademacher Average): For a class $\mathcal{F}$ of functions, the Rademacher Average is defined as

$$\mathcal{R}(\mathcal{F}) = \mathbb{E}_S(\hat{\mathcal{R}}_S(\mathcal{F})) = \mathbb{E}_{\boldsymbol{\sigma}, S} \sup_{f \in \mathcal{F}} (\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i))$$

Rademacher Average

In $\mathcal{R}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma},S} \sup_{f \in \mathcal{F}}(\frac{1}{m}\sum_{i=1}^{m} \sigma_i f(x_i))$ we define $f \in \{-1, 1\}$. Generally, f can be any function of $\mathbb{N} \to \mathbb{R}$, this gives the general definition of Rademacher Average (See reference [2]).

Define $g(x_i) = \mathbf{1}_{f(x_i) \neq y_i} \in \{0, 1\}$, and the set of all gs is a function class $\mathcal{G}$. Define $\mathcal{R}(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma},S} \sup_{g \in \mathcal{G}}(\frac{1}{m}\sum_{i=1}^{m} \sigma_i g(x_i))$. Then we have the following relationship between $\mathcal{R}(\mathcal{F})$ and $\mathcal{R}(\mathcal{G})$:

$$
\begin{aligned}
\mathcal{R}(\mathcal{G}) &= \mathbb{E}_{\boldsymbol{\sigma},S} \sup_{g \in \mathcal{G}}(\frac{1}{m}\sum_{i=1}^{m} \sigma_i g(x_i)) = \mathbb{E}_{\boldsymbol{\sigma},S} \sup_{f \in \mathcal{F}}(\frac{1}{m}\sum_{i=1}^{m} \sigma_i \mathbf{1}_{f(x_i) \neq y_i}) \\
&= \mathbb{E}_{\boldsymbol{\sigma},S} \sup_{f \in \mathcal{F}}(\frac{1}{m}\sum_{i=1}^{m} \sigma_i \frac{1 - y_i f(x_i)}{2}) \\
&= \mathbb{E}_{\boldsymbol{\sigma},S} \left[\frac{1}{2m}\sum_{i=1}^{m} \sigma_i + \sup_{f \in \mathcal{F}}(\frac{1}{2m}\sum_{i=1}^{m}(-\sigma_i y_i) f(x_i)) \right] \\
&= \mathbb{E}_{\boldsymbol{\sigma},S} \sup_{f \in \mathcal{F}}(\frac{1}{2m}\sum_{i=1}^{m} \sigma_i f(x_i)) = \frac{1}{2}\mathcal{R}(\mathcal{F})
\end{aligned}
$$

Rademacher Bound Theorem

Theorem (Rademacher Bound): Let $g(x_i) = \mathbf{1}_{f(x_i) \neq y_i} \in \{0, 1\}$, $\mathcal{R}(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{g \in \mathcal{G}} (\frac{1}{m} \sum_{i=1}^{m} \sigma_i g(x_i))$ be the Rademacher Average of $\mathcal{F}$, then

$$P\left[\sup_{f \in \mathcal{F}} (R^{\text{true}}(f) - R^{\text{emp}}(f)) \leq 2\mathcal{R}(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}\right] \geq 1 - \delta$$

Remember that $R^{\text{true}}(f)$ is the true risk, $R^{\text{emp}}(f) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{f(x_i) \neq y_i}$ is the empirical risk and $R^{\text{true}}(f) = E(R^{\text{emp}}(f))$.

To prove this bound, I will first introduce some basics about conditional expectation and supremum, then talk about McDiarmid's Inequality, Symmetrization and Lemma of Rademacher Average. Finally I will put these components together.

Basics: Conditional Expectation

- Conditional Expectation: Suppose we have a bivariate random vector (X, Y) with joint PMF $P(X, Y)$. We call $E_X(X|Y = y)$ the expectation of X given that $Y = y$. Sometimes we just simplify the notation as $E(X|Y)$. Formally,

$$E(X|Y) = \sum_x xP(X = x|Y = y) = \sum_x x\frac{P(X = x, Y = y)}{P(Y = y)}$$

Note that $E_X(X|Y = y)$ is a function of y.

Properties:
(1) If Y is independent of X, then $E(X|Y) = E(X)$.
(2) Let $f(Z)$ be a function of a random variable Z, then $E(f(Z)|Z) = f(Z)$.

- Law of Total Expectation:

$$E(X) = E_Y(E_X(X|Y))$$

Basics: Supremum

- Let $t \in \mathcal{T}$ be a variable, f and g be two functions of t, then

$$\sup\{f(t) + g(t)\} \leq \sup\{f(t)\} + \sup\{g(t)\}$$

Proof: Since $f(t) \leq \sup\{f(t)\}$ and $g(t) \leq \sup\{g(t)\}$, we have $f(t) + g(t) \leq \sup\{f(t)\} + \sup\{g(t)\}$ for any t, thus $\sup\{f(t) + g(t)\} \leq \sup\{f(t)\} + \sup\{g(t)\}$.

- Switching Supremum and Expectation: Let $f(X, y)$ be any function of two variables X and y, and $y \in \mathcal{Y}$, then

$$\sup_{y \in \mathcal{Y}} \mathbb{E} f(X, y) \leq \mathbb{E} \sup_{y \in \mathcal{Y}} f(X, y)$$

Proof: Since $f(X, y) \leq \sup_{y \in \mathcal{Y}} f(X, y)$ for any X, we take expectation with respect to X on both sides, then $\mathbb{E} f(X, y) \leq \mathbb{E} \sup_{y \in \mathcal{Y}} f(X, y)$. Since this inequality holds for any y, we have $\sup_{y \in \mathcal{Y}} \mathbb{E} f(X, y) \leq \mathbb{E} \sup_{y \in \mathcal{Y}} f(X, y)$.

McDiarmid's Inequality

McDiarmid's Inequality: Define the function $F : \mathcal{X}^m \to \mathbb{R}$. Let $(x_1, x_2, ..., x_m) \in \mathcal{X}^m$. Suppose for any i, replacing x_i by any x_i', the following equality holds:

$$\sup_{x_i' \in \mathcal{X}} |F(x_1, ..., x_i, ..., x_m) - F(x_1, ..., x_i', ..., x_m)| \leq c$$

where c is a constant. Then for all $\epsilon > 0$,

$$P(|F - \mathbb{E}(F)| \geq \epsilon) \leq 2 \exp\left(-\frac{2\epsilon^2}{mc^2}\right)$$

Proof: Use the marginal sequence, suppose

$$V_i = \mathbb{E}(F|x_1, x_2, ..., x_i) - \mathbb{E}(F|x_1, x_2, ..., x_{i-1})$$

for $i = 1, 2, ..., m$. Then

$$\mathbb{E}(V_i) = \mathbb{E}(\mathbb{E}(F|x_1, x_2, ..., x_i)) - \mathbb{E}(\mathbb{E}(F|x_1, x_2, ..., x_{i-1})) = \mathbb{E}(F) - \mathbb{E}(F) = 0$$

$$\sum_{i=1}^{m} V_i = \mathbb{E}(F|x_1, x_2, ..., x_m) - \mathbb{E}(F) = F - \mathbb{E}(F)$$

$$
\begin{aligned}
\sup V_i - \inf V_i = &\left[\sup_x \mathbb{E}(F|x_1, x_2, ..., x_{i-1}, x) - \mathbb{E}(F|x_1, x_2, ..., x_{i-1})\right] - \\
&\left[\inf_x \mathbb{E}(F|x_1, x_2, ..., x_{i-1}, x) - \mathbb{E}(F|x_1, x_2, ..., x_{i-1})\right] \\
= &\sup_x \mathbb{E}(F|x_1, x_2, ..., x_{i-1}, x) - \inf_x \mathbb{E}(F|x_1, x_2, ..., x_{i-1}, x) \\
= &\sup_{x_u, x_l} \mathbb{E}(F|x_1, x_2, ..., x_{i-1}, x_u) - \mathbb{E}(F|x_1, x_2, ..., x_{i-1}, x_l) \\
= &\sup_{x_u, x_l} \sum_{x_{i+1}, \ldots, x_m} [F(x_1, ..., x_{i-1}, x_u, x_{i+1}, ..., x_m|x_1, ..., x_{i-1}) - \\
&F(x_1, ..., x_{i-1}, x_l, x_{i+1}, ..., x_m|x_1, ..., x_{i-1})]P(x_{i+1}, ..., x_m) \\
\leq &\, c \sum_{x_{i+1}, \ldots, x_m} P(x_{i+1}, ..., x_m) = c
\end{aligned}
$$

By Azuma's Inequality (see Appendix),

$$
P(|F - \mathbb{E}(F)| \geq \epsilon) = P(|\sum_{i=1}^{m} V_i| \geq \epsilon) \leq 2 \exp\left(-\frac{2\epsilon^2}{mc^2}\right)
$$

Symmetrization

Let $g(x_i) = \mathbf{1}_{f(x_i) \neq y_i} \in \{0, 1\}$. Since $R^{\text{emp}}(f) = \frac{1}{m} \sum_{j=1}^{m} g(x_j)$, we can consider $R^{\text{emp}}(f)$ as a function of $g(x_1), g(x_2), ..., g(x_m)$. Define $R^{\text{emp,i}}(f) = \frac{1}{m}(\sum_{j=1, j \neq i}^{m} g(x_j)$ $+ g(x_i'))$, that is, $g(x_i)$ is replaced by $g(x_i')$. Then the following inequality holds:

$$|\sup_{f \in \mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) - \sup_{f \in \mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp,i}}(f))| \leq \sup_{f \in \mathcal{F}} |R^{\text{emp,i}}(f) - R^{\text{emp}}(f)| \quad (*)$$

To prove this, suppose f^* achieves $\sup_{f \in \mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f))$, and $\sup_{f \in \mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) > \sup_{f \in \mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp,i}}(f))$, then

$$\begin{aligned}
&\sup_{f \in \mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) - \sup_{f \in \mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp,i}}(f)) \qquad (\diamond) \\
&= (R^{\text{true}}(f^*) - R^{\text{emp}}(f^*)) - \sup_{f \in \mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp,i}}(f)) \\
&\leq (R^{\text{true}}(f^*) - R^{\text{emp}}(f^*)) - (R^{\text{true}}(f^*) - R^{\text{emp,i}}(f^*)) \\
&= R^{\text{emp,i}}(f^*) - R^{\text{emp}}(f^*) \leq |R^{\text{emp,i}}(f^*) - R^{\text{emp}}(f^*)| \leq \sup_{f \in \mathcal{F}} |R^{\text{emp,i}}(f) - R^{\text{emp}}(f)|
\end{aligned}$$

When $\sup_{f \in \mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) \leq \sup_{f \in \mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp,i}}(f))$, switching the first and second terms in $(\diamond)$ and we can get the same result, thus the inequality $(*)$ holds.

Since $|R^{\text{emp}}(f) - R^{\text{emp,i}}(f)| = \frac{1}{m}|Z_i - Z_i'| \leq \frac{1}{m}$, we have

$$|\sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) - \sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp,i}}(f))| \leq \frac{1}{m}$$

This satisfies the condition of McDiarmid's inequality with $c = \frac{1}{m}$, thus

$$P\left[|\sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) - \mathbb{E}_S \sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f))| \geq \epsilon\right] \leq 2\exp\left(-2m\epsilon^2\right)$$

Since the McDiarmid's Bound is symmetric, for one-sided bound, we have

$$P\left[\sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) - \mathbb{E}_S \sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) \geq \epsilon\right] \leq \exp\left(-2m\epsilon^2\right)$$

Lemma of Rademacher Average

Lemma: Let $S = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ be a set of training samples, and $R^{\mathsf{emp}}(f) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{f(x_i)\neq y_i} = \frac{1}{m}\sum_{i=1}^{m} g(x_i)$, then

$$\mathbb{E}_S \sup_{f\in\mathcal{F}}[R^{\mathsf{true}}(f) - R^{\mathsf{emp}}(f)] \le 2\mathbb{E}_{\boldsymbol{\sigma},S} \sup_{f\in\mathcal{F}}(\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(x_i)) = 2\mathcal{R}(\mathcal{G})$$

Proof: Suppose we have a set of ghost samples $S' = \{(x'_1, y'_1), (x'_2, y'_2), ..., (x'_m, y'_m)\}$ and the empirical risk on the ghost samples is $R^{\mathsf{emp}\prime}(f) = \frac{1}{m}\sum_{i=1}^{m} g(x'_i)$. Since $R^{\mathsf{true}}(f) = \mathbb{E}_{S'}[R^{\mathsf{emp}\prime}(f)|S]$ and $R^{\mathsf{emp}}(f) = \mathbb{E}_{S'}[R^{\mathsf{emp}}(f)|S]$. Then,

$$\begin{aligned}
\mathbb{E}_S \sup_{f\in\mathcal{F}}[R^{\mathsf{true}}(f) - R^{\mathsf{emp}}(f)] &= \mathbb{E}_S \sup_{f\in\mathcal{F}}[\mathbb{E}_{S'}[R^{\mathsf{emp}\prime}(f)|S] - \mathbb{E}_{S'}[R^{\mathsf{emp}}(f)|S]] \\
&= \mathbb{E}_S \sup_{f\in\mathcal{F}}[\mathbb{E}_{S'}[(R^{\mathsf{emp}\prime}(f) - R^{\mathsf{emp}}(f))|S]] \\
&\le \mathbb{E}_S\mathbb{E}_{S'} \sup_{f\in\mathcal{F}}[(R^{\mathsf{emp}\prime}(f) - R^{\mathsf{emp}}(f))|S] \qquad (\diamond) \\
&= \mathbb{E}_{S,S'}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}(g(x'_i) - g(x_i))\right]
\end{aligned}$$

In $(\diamond)$ we use the property that $\sup_{y\in\mathcal{Y}}\mathbb{E}f(X,y) \le \mathbb{E}\sup_{y\in\mathcal{Y}} f(X,y)$.

Now we introduce a set of Rademacher variables $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, ..., \sigma_m\}$. Each σ_i satisfies $P(\sigma_i = -1) = P(\sigma_i = 1) = \frac{1}{2}$.

$$
\begin{aligned}
\mathbb{E}_S \sup_{f \in \mathcal{F}} [R^{\textsf{true}}(f) - R^{\textsf{emp}}(f)] &\leq \mathbb{E}_{S,S'} \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} (g(x_i') - g(x_i)) \right] \\
&= \mathbb{E}_{\boldsymbol{\sigma},S,S'} \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i (g(x_i') - g(x_i)) \right] \qquad (*) \\
&\leq \mathbb{E}_{\boldsymbol{\sigma},S'} \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(x_i') \right] + \mathbb{E}_{\boldsymbol{\sigma},S} \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} -\sigma_i g(x_i) \right] \qquad (\star) \\
&= 2\mathbb{E}_{\boldsymbol{\sigma},S} \sup_{g \in \mathcal{G}} (\frac{1}{m} \sum_{i=1}^{m} \sigma_i g(x_i)) = 2\mathcal{R}(\mathcal{G})
\end{aligned}
$$

The equality $(*)$ uses the fact that, since $g(x_i')$ and $g(x_i)$ are interchangeable and $\sigma_i = -1$ or 1, multiplying σ_i will not change the distribution of $f(x_i') - f(x_i)$. The inequality $(\star)$ uses the fact $\sup\{f(t) + g(t)\} \leq \sup\{f(t)\} + \sup\{g(t)\}$.

Putting All Components Together

By Symmertrization we have the following one-sided bound:

$$P\left[\sup_{f\in\mathcal{F}}(R^{\textsf{true}}(f)-R^{\textsf{emp}}(f))-\mathbb{E}\sup_{f\in\mathcal{F}}(R^{\textsf{true}}(f)-R^{\textsf{emp}}(f))\geq\epsilon\right]\leq\exp\left(-2m\epsilon^2\right)$$

Let $\delta=\exp\left(-2m\epsilon^2\right)$, so that $\epsilon=\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$, thus

$$P\left[\sup_{f\in\mathcal{F}}(R^{\textsf{true}}(f)-R^{\textsf{emp}}(f))-\mathbb{E}\sup_{f\in\mathcal{F}}(R^{\textsf{true}}(f)-R^{\textsf{emp}}(f))\geq\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right]\leq\delta$$

$$P\left[\sup_{f\in\mathcal{F}}(R^{\textsf{true}}(f)-R^{\textsf{emp}}(f))-\mathbb{E}\sup_{f\in\mathcal{F}}(R^{\textsf{true}}(f)-R^{\textsf{emp}}(f))\leq\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right]\geq 1-\delta$$

$$P\left[\sup_{f\in\mathcal{F}}(R^{\textsf{true}}(f)-R^{\textsf{emp}}(f))\leq\mathbb{E}\sup_{f\in\mathcal{F}}(R^{\textsf{true}}(f)-R^{\textsf{emp}}(f))+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right]\geq 1-\delta$$

By Lemma of Rademacher Average, $\mathbb{E}\sup_{f\in\mathcal{F}}(R^{\textsf{true}}(f)-R^{\textsf{emp}}(f))\leq 2\mathcal{R}(\mathcal{G})$, therefore

$$P\left[\sup_{f\in\mathcal{F}}(R^{\textsf{true}}(f)-R^{\textsf{emp}}(f))\leq 2\mathcal{R}(\mathcal{G})+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right]\geq 1-\delta$$

The bound of Empirical Rademacher Average

We can also bound $\sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f))$ by Empirical Rademacher Average. Since $\hat{\mathcal{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_{g\in\mathcal{G}}(\frac{1}{m}\sum_{i=1}^m \sigma_i g(x_i))$, changing one x_i will change $\hat{\mathcal{R}}_S(\mathcal{G})$ by at most $\frac{1}{m}$. Then by McDiarmid's Inequality

$$
\begin{aligned}
P\left[\left|\hat{\mathcal{R}}_S(\mathcal{G}) - \mathcal{R}(\mathcal{G})\right| > \epsilon\right] &\le 2\exp\left(-2m\epsilon^2\right) \\
P\left[\hat{\mathcal{R}}_S(\mathcal{G}) - \mathcal{R}(\mathcal{G}) > \epsilon\right] &\le \exp\left(-2m\epsilon^2\right) \qquad (1)
\end{aligned}
$$

Since

$$
P\left[\sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) \le 2\mathcal{R}(\mathcal{G}) + \epsilon\right] \ge 1 - \exp\left(-2m\epsilon^2\right)
$$

$$
P\left[\sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) \ge 2\mathcal{R}(\mathcal{G}) + \epsilon\right] \le \exp\left(-2m\epsilon^2\right) \qquad (2)
$$

By (1) and (2), we have

$$
\begin{aligned}
&P\left[\sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) \ge 2\hat{\mathcal{R}}_S(\mathcal{G}) + 3\epsilon\right] \\
&\le P\left[\sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) \ge 2\mathcal{R}(\mathcal{G}) + \epsilon \text{ or } \hat{\mathcal{R}}_S(\mathcal{G}) - \mathcal{R}(\mathcal{G}) \ge \epsilon\right] \\
&\le P\left[\sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) \ge 2\mathcal{R}(\mathcal{G}) + \epsilon\right] + P\left[\hat{\mathcal{R}}_S(\mathcal{G}) - \mathcal{R}(\mathcal{G}) \ge \epsilon\right] \le 2\exp\left(-2m\epsilon^2\right)
\end{aligned}
$$

Let $\delta = 2\exp\left(-2m\epsilon^2\right)$, so that $\epsilon = \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$, thus we have the following Empirical Rademacher Average bound:

$$P\left[\sup_{f\in\mathcal{F}}(R^{\mathsf{true}}(f) - R^{\mathsf{emp}}(f)) \geq 2\hat{\mathcal{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \leq \delta$$

$$P\left[\sup_{f\in\mathcal{F}}(R^{\mathsf{true}}(f) - R^{\mathsf{emp}}(f)) \leq 2\hat{\mathcal{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right] \geq 1-\delta$$

Massart's Lemma

Since we have

$$P\left[\sup_{f\in\mathcal{F}}(R^{\text{true}}(f) - R^{\text{emp}}(f)) \leq 2\mathcal{R}(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right] \geq 1-\delta$$

The question is, how can we calculate $\mathcal{R}(\mathcal{G})$? The following lemma gives an upperbound:

Massart's Lemma: Let $A \subset \mathbb{R}^m$ be a finite set of points and $a = \{a_1, a_2, ..., a_m\} \in A$. Let $r = \sup_{a\in A}\|a\|_2$, then

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{a\in A}\sum_{i=1}^{m}\sigma_i a_i\right] \leq r\sqrt{2\log|A|}$$

Proof: Let $t > 0$ be a number to be chosen later.

$$\exp\left(t\mathbb{E}_{\boldsymbol{\sigma}}\sup_{a\in A}\sum_{i=1}^{m}\sigma_i a_i\right) \leq \mathbb{E}_{\boldsymbol{\sigma}}\exp\left(t\sup_{a\in A}\sum_{i=1}^{m}\sigma_i a_i\right) = \mathbb{E}_{\boldsymbol{\sigma}}\sup_{a\in A}\exp\left(\sum_{i=1}^{m}t\sigma_i a_i\right)$$

$$= \mathbb{E}_{\boldsymbol{\sigma}}\sup_{a\in A}\left(\prod_{i=1}^{m}\exp(t\sigma_i a_i)\right) \leq \mathbb{E}_{\boldsymbol{\sigma}}\sum_{a\in A}\prod_{i=1}^{m}\exp(t\sigma_i a_i) = \sum_{a\in A}\prod_{i=1}^{m}\mathbb{E}_{\boldsymbol{\sigma}}\exp(t\sigma_i a_i)$$

$$\leq \sum_{a\in A}\prod_{i=1}^{m}\exp\left(\frac{t^2a_i^2}{2}\right) = \sum_{a\in A}\exp\left(\frac{t^2(\sum_{i=1}^{m}a_i^2)}{2}\right) \leq |A|\exp\left(\frac{t^2r^2}{2}\right)$$

The $\leq$ uses the Jenson's Inequality. The $\leq$ uses Hoeffding's Lemma, given that $-|a_i| \leq \sigma_i a_i \leq |a_i|$.

Therefore, taking $\log$ on both sides, we get

$$\mathbb{E}_{\boldsymbol{\sigma}} \sup_{a \in A} \sum_{i=1}^{m} \sigma_i a_i \leq \inf_{t>0} \left(\frac{\log |A|}{t} + \frac{tr^2}{2} \right) = r\sqrt{2 \log |A|}$$

Let $A = \mathcal{F}$, $|A| = S_m(\mathcal{F})$, $a_i = f(x_i) \in \{-1, 1\}$, then $r = \sqrt{m}$. Therefore

$$\mathcal{R}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma},S} \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \leq \frac{r}{m} \sqrt{2 \log S_m(\mathcal{F})} = \sqrt{\frac{2 \log S_m(\mathcal{F})}{m}}$$

By Sauer's Lemma, $S_m(\mathcal{F}) \leq (\frac{em}{h})^h$, where h is the VC dimension of $\mathcal{F}$, then

$$\mathcal{R}(\mathcal{G}) = \frac{1}{2}\mathcal{R}(\mathcal{F}) \leq \frac{1}{2}\sqrt{\frac{2 \log S_m(\mathcal{F})}{m}} \leq \frac{1}{2}\sqrt{\frac{2h \log \frac{em}{h}}{m}}$$

Thus the Rademacher Bound becomes

$$P\left[\sup_{f \in \mathcal{F}} (R^{\text{true}}(f) - R^{\text{emp}}(f)) \leq \sqrt{\frac{2h \log \frac{em}{h}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \right] \geq 1 - \delta$$

Comparing the Rademacher Bound and the Growth Function Bound, both bounds grow like $O(\sqrt{\frac{h\log(em/h)}{m}})$ with respect to m and h, so the Rademacher Bound will not be looser than the Growth Function Bound.

Actually Massart's Lemma gives a loose bound for $\mathcal{R}(\mathcal{G})$. Sometimes we bound $\mathcal{R}(\mathcal{G})$ by a constant instead of Massart's Lemma. For example, since $g(x_i) \in \{0, 1\}$

$$\mathcal{R}(\mathcal{G}) = \frac{1}{2}\mathcal{R}(\mathcal{F}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma},S} \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i) \leq \frac{1}{2}$$

$\mathcal{R}(\mathcal{G})$ reaches its maximum when given any dataset S, there always exists an $f \in \mathcal{F}$ that perfectly classifies it. This gives another upperbound for $\mathcal{R}(\mathcal{G})$.

Contents

Covering Number

Recall that we have a dataset $S = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$, and a function class $\mathcal{F}$. The functions $f \in \mathcal{F}$ are binary-valued, i.e., $f(x_i) \in \{-1, 1\}$. The Growth Function partitions $\mathcal{F}$ into finite number of groups because the cardinality of the set $\{-1, 1\}$ is finite. Since the cardinality of $\{-1, 1\}$ is 2 and we have m samples, we can partition $\mathcal{F}$ into at most 2^m groups.

However, when f is continuous, the candinarity of the value set will be infinite, thus the Growth Function will not work. To address this, we introduce another way to measure the capacity of $\mathcal{F}$ – Covering Number.

Let $(\mathcal{F}, d)$ be a metric space, and we define the distance of a function $f \in \mathcal{F}$ by L_p norm of m points:

$$\|f\|_{L_p(m)} = \left(\frac{1}{m} \sum_{i=1}^{m} |f(x_i)|^p \right)^{1/p}$$

The distance of two functions f and f' can be defined as

$$\|f - f'\|_{L_p(m)} = \left(\frac{1}{m} \sum_{i=1}^{m} |f(x_i) - f'(x_i)|^p \right)^{1/p}$$

Let's use L_1 norm to define the distance, then

$$\|f - f'\|_{L_1(m)} = \left(\frac{1}{m} \sum_{i=1}^{m} |f(x_i) - f'(x_i)| \right)$$

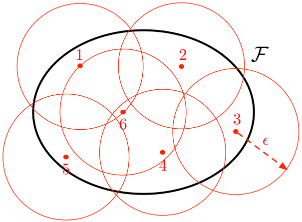
Covering Number

Denote $B(f_j, \epsilon)$ as a ball of radius ϵ around f_j, using the metric $L_p(m)$. $B(f_j, \epsilon)$ contains all the function in $\mathcal{F}$ that are within distance ϵ of f_j, i.e., $B(f_j, \epsilon) = \{f' : \|f - f'\|_{L_p(m)} < \epsilon\}$. Here $B(f_j, \epsilon)$ is considered as one group of functions, and we would like to know how many groups can cover $\mathcal{F}$. We say that $f_1, f_2, ..., f_N$ covers $\mathcal{F}$ at radius ϵ if:

$$\mathcal{F} \subset \bigcup_{j=1}^{N} B(f_j, \epsilon)$$

Definition (Covering Number): The covering number of $\mathcal{F}$ at radius $\mathcal{F}$ with respect to $L_p(m)$, denoted by $N(\mathcal{F}, \epsilon, L_p(m))$, is the minimum N such that $f_1, f_2, ..., f_N$ covers $\mathcal{F}$ at radius ϵ.

If the covering number is finite, we can approximately represent $\mathcal{F}$ by a finite set of functions that cover $\mathcal{F}$. This is another a finite measure for the capacity of $\mathcal{F}$.



Discretization Theorem

We can bound the empirical Rademacher Average with the following theorem:

Theorem (Discretication): Let $-1 \leq f \leq 1$ for all $f \in \mathcal{F}$, then for any $S = \{x_1, x_2, ..., x_m\}$ we have

$$\hat{\mathcal{R}}_S(\mathcal{F}) \leq \inf_{\epsilon \geq 0} \{\epsilon + \sqrt{\frac{2\log(2N(\mathcal{F}, \epsilon, L_1(m)))}{m}}\}$$

Proof: Fix $S = \{x_1, x_2, ..., x_m\}$ and $\epsilon > 0$. Let V be the minimal set of ϵ-balls that covers $\mathcal{F}$, thus $|V| = N(\mathcal{F}, \epsilon, L_1(m))$. For any $f \in \mathcal{F}$, define $f' \in V$ such that $\|f - f'\|_{L_1(m)} < \epsilon$. Then

$$\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{F}) &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} (\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i)) \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} (\frac{1}{m} \sum_{i=1}^{m} \sigma_i (f(x_i) - f'(x_i))) + \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} (\frac{1}{m} \sum_{i=1}^{m} \sigma_i f'(x_i)) \\
&\leq \epsilon + \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{F}} (\frac{1}{m} \sum_{i=1}^{m} \sigma_i f'(x_i)) \\
&\leq \epsilon + \sqrt{\frac{2\log(2|V|)}{m}} \qquad \text{[Massart's Lemma]} \\
&\leq \epsilon + \sqrt{\frac{2\log(2N(\mathcal{F}, \epsilon, L_1(m)))}{m}}
\end{aligned}$$

Note that this inequality holds for any $\epsilon \geq 0$, so we can add an $\inf_{\epsilon \geq 0}$ on the RHS.

Covering Theorem

The following theorem gives a generalization error bound using Covering Number:

Theorem (Covering): Let $R^{\text{emp}}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$ and $R^{\text{true}}(f) = \mathbb{E}[R^{\text{emp}}(f)]$. Suppose there exists $M > 0$ such that $|f(x_i) - y_i| \leq M$ for any x_i. Then for any $\epsilon > 0$,

$$P\left[\sup_{f\in\mathcal{F}} \left|R^{\text{true}}(f) - R^{\text{emp}}(f)\right| \geq \epsilon\right] \leq 2N(\mathcal{F}, \frac{\epsilon}{8M}, L_\infty(m)) \exp\left(-\frac{m\epsilon^2}{2M^4}\right)$$

Proof:
Let $L_S(f) = R^{\text{true}}(f) - R^{\text{emp}}(f)$, we show that for all $f_1, f_2 \in \mathcal{F}$ and any dataset S, the following inequality holds:

$$|L_S(f_1) - L_S(f_2)| \leq 4M\|f_1 - f_2\|_{L_\infty(m)}$$

This is because

$$|L_S(f_1) - L_S(f_2)| \leq |R^{\text{true}}(f_1) - R^{\text{true}}(f_2)| + |R^{\text{emp}}(f_1) - R^{\text{emp}}(f_2)|$$
$$= |\mathbb{E}_{x,y}[(f_1(x) - y)^2 - (f_2(x) - y)^2]| + \frac{1}{m}\sum_{i=1}^{m}[(f_1(x_i) - y_i)^2 - (f_2(x_i) - y_i)^2]$$

Since

$$(f_1(x) - y)^2 - (f_2(x) - y)^2 = (f_1(x) - f_2(x))[(f_1(x) - y) + (f_2(x) - y)] \leq \|f_1 - f_2\|_{L_\infty(m)} 2M$$

We have $|L_S(f_1) - L_S(f_2)| \leq 4M\|f_1 - f_2\|_{L_\infty(m)}$.

Assume $\mathcal{F}$ can be covered by $B_1, B_2, ..., B_k$ such that $\mathcal{F} \subset \bigcup_{i=1}^{k} B_i$, then

$$P\left[\sup_{f\in\mathcal{F}} \left|R^{\text{true}}(f) - R^{\text{emp}}(f)\right| \geq \epsilon\right] = P\left[\bigcup_{i=1}^{k} \sup_{f\in B_i} \left|R^{\text{true}}(f) - R^{\text{emp}}(f)\right| \geq \epsilon\right]$$
$$\leq \sum_{i=1}^{k} P\left[\sup_{f\in B_i} \left|R^{\text{true}}(f) - R^{\text{emp}}(f)\right| \geq \epsilon\right] \qquad (*)$$

Let f_i be the center of the ball B_i and the radius of B_i is $\frac{\epsilon}{8M}$, then for all $f \in B_i$, we have $\|f - f_i\|_{L_\infty(m)} \leq \frac{\epsilon}{8M}$, thus

$$|L_S(f) - L_S(f_i)| \leq 4M\|f - f_i\|_{L_\infty(m)} \leq \frac{\epsilon}{2}$$

Therefore, let the events $A : |L_S(f)| \geq \epsilon$ and $B : |L_S(f_i)| \geq \frac{\epsilon}{2}$, we must have $A \Rightarrow B$. Thus for any $f \in B_i$

$$P\left[\left|R^{\text{true}}(f) - R^{\text{emp}}(f)\right| \geq \epsilon\right] \leq P\left[\left|R^{\text{true}}(f_i) - R^{\text{emp}}(f_i)\right| \geq \frac{\epsilon}{2}\right]$$

Let A, B, C be three events. If $A \Rightarrow C$ and $B \Rightarrow C$, then $A \cup B \Rightarrow C$, thus $P(A \cup B) \leq P(C)$. We have

$$P\left[\sup_{f\in B_i} \left|R^{\text{true}}(f) - R^{\text{emp}}(f)\right| \geq \epsilon\right] = P\left[\bigcup_{f\in B_i} \left|R^{\text{true}}(f) - R^{\text{emp}}(f)\right| \geq \epsilon\right]$$
$$\leq P\left[\left|R^{\text{true}}(f_i) - R^{\text{emp}}(f_i)\right| \geq \frac{\epsilon}{2}\right]$$

Since $R^{\text{emp}}(f_j) = \frac{1}{m}\sum_{i=1}^{m}(f_j(x_i) - y_i)^2 \leq M^2$, use Hoeffding Bound, we have

$$P\left[\left|R^{\text{true}}(f_i) - R^{\text{emp}}(f_i)\right| \geq \frac{\epsilon}{2}\right] \leq 2\exp\left(-\frac{2m(\epsilon/2)^2}{(M^2)^2}\right) = 2\exp\left(-\frac{m\epsilon^2}{2M^4}\right)$$

Since k is the minimum number of the balls B_i, and each B_i is with radius $\frac{\epsilon}{8M}$, we have $k = N(\mathcal{F}, \frac{\epsilon}{8M}, L_\infty(m))$.

Putting all things together, we get

$$P\left[\sup_{f\in\mathcal{F}}\left|R^{\text{true}}(f) - R^{\text{emp}}(f)\right| \geq \epsilon\right] \leq 2N(\mathcal{F}, \frac{\epsilon}{8M}, L_\infty(m))\exp\left(-\frac{m\epsilon^2}{2M^4}\right)$$

References

[1] Prediction: Machine Learning And Statistics. Lecture 14. https://ocw.mit.edu/courses/15-097-prediction-machine-learning-and-statistics-spring-2012/pages/lecture-notes/

[2] Theoretical Machine Learning. Lecture 9. https://www.cs.princeton.edu/courses/archive/spring13/cos511/scribe_notes/0305.pdf

[3] Machine Learning Theory: Rademacher Complexity https://www.cs.cmu.edu/~ninamf/ML11/lect1117.pdf

[4] Rademacher Complexity and Massart's Lemma https://web.eecs.umich.edu/~jabernet/eecs598course/fall2015/web/notes/lec13_102715.pdf

[5] Mathematics of Machine Learning Lecture 6: Covering Number https://ocw.mit.edu/courses/18-657-mathematics-of-machine-learning-fall-2015/resources/mit18_657f15_l6/

[6] https://cs.nyu.edu/~mohri/ml13/sol2.pdf

References

[7] Notes 20 : Azuma's inequality
https://people.math.wisc.edu/~roch/grad-prob/gradprob-notes20.pdf

[8] https://people.eecs.berkeley.edu/~sinclair/cs271/n18.pdf

[9] Concentration Inequalities: Hoeffding and McDiarmid
https://people.eecs.berkeley.edu/~bartlett/courses/281b-sp08/13.pdf

Appendix: Azuma's Inequality

Azuma's Inequality is a generalization of Hoeffding's Inequality.

Azuma's Inequality: Suppose $F : \mathcal{X}^m \to \mathbb{R}$ is a function of variables $x_1, x_2, ..., x_m$. Let $Z_i = \mathbb{E}(F|x_1, x_2, ..., x_i)$ for $i \geq 1$ and $Z_0 = \mathbb{E}(F)$. Let $V_i = Z_i - Z_{i-1}$ be the marginal sequence, and assume $\sup V_i - \inf V_i \leq c$ for $i = 1, 2, ..., m$, where c is a constant. Then

$$P\left[Z_m - Z_0 \geq \epsilon\right] = P\left[\sum_{i=1}^{m} V_i \geq \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{mc^2}\right)$$

Proof: Let $t > 0$ be a value to be determined

$$P\left[\sum_{i=1}^{m} V_i \geq \epsilon\right] \leq e^{-t\epsilon}\mathbb{E}\left[e^{t\sum_{i=1}^{m} V_i}\right] \qquad \text{[Chernoff Bound]}$$
$$= e^{-t\epsilon}\mathbb{E}\left[\mathbb{E}\left[e^{t\sum_{i=1}^{m} V_i}\middle| x_1, x_2, ..., x_{m-1}\right]\right] \qquad \text{[Law of the total expectation]}$$
$$= e^{-t\epsilon}\mathbb{E}\left[e^{t\sum_{i=1}^{m-1} V_i}\mathbb{E}\left[e^{tV_m}\middle| x_1, x_2, ..., x_{m-1}\right]\right]$$

The last equality is because $\sum_{i=1}^{m-1} V_i$ can be considered as a function of $x_1, x_2, ..., x_{m-1}$. Let h be a function of X, we have

$$\mathbb{E}_X[\mathbb{E}_Y[h(X)Y|X]] = \mathbb{E}_X[h(X)\mathbb{E}_Y[Y|X]]$$

Appendix: Azuma's Inequality

Recall the Hoeffding Lemma: Let X be any random variable such that $a \leq X \leq b$, for all $t \in \mathbb{R}$, $\mathbb{E}[e^{tX}] \leq \exp(\frac{t^2(b-a)^2}{8})$. Therefore,

$$\mathbb{E}\left[e^{tV_m} \middle| x_1, x_2, ..., x_{m-1}\right] \leq e^{\frac{t^2c^2}{8}}$$

Induct the rest part $e^{t\sum_{i=1}^{m-1} V_i}$ by the following formula:

$$\mathbb{E}\left[e^{t\sum_{i=1}^{k} V_i}\right] = \mathbb{E}\left[e^{t\sum_{i=1}^{k-1} V_i}\mathbb{E}\left[e^{tV_k} \middle| x_1, x_2, ..., x_{k-1}\right]\right]$$

for $k = m-1, m-2, ..., 1$, we get

$$P\left[\sum_{i=1}^{m} V_i \geq \epsilon\right] \leq e^{-t\epsilon}e^{\frac{mt^2c^2}{8}} = \exp\left(-t\epsilon + \frac{mt^2c^2}{8}\right)$$

The above inequality holds for any t. Let $t = \frac{4\epsilon}{mc^2}$, we get $\inf_t(-t\epsilon + \frac{mt^2c^2}{8}) = -\frac{2\epsilon}{mc^2}$, thus

$$P\left[\sum_{i=1}^{m} V_i \geq \epsilon\right] \leq \exp\left(-\frac{2\epsilon}{mc^2}\right)$$

Appendix: Azuma's Inequality

Since the Chernoff Bound is symmetric, for two-sided bound, we have

$$P\left[\left|\sum_{i=1}^{m} V_i\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2\epsilon^2}{mc^2}\right)$$

Note that Azuma's Inequality does not assume $V_1, V_2, ..., V_m$ to be independent. In fact $V_1, V_2, ..., V_m$ are can be dependent (see [8]), we can not apply Hoeffding's Inequality to them. That is the reason we use Azuma's Inequality instead.