Indian Institute of Technology, Kanpur
Statistical and Algorithmic Learning Theory

Instructors: Purushottam Kar
Authors: Gurpreet Singh
Email: guggu@iitk.ac.in

SCRIBE

7

The McDiarmid Inequality

1. Recap

In the last scribe, we discussed concentration bounds for infinite hypothesis classes which can be represented by an ϵ-cover, using Hoeffding's Inequality. However, the concentration bound derived was dependent on the dimensionality of the feature space. Thus, the concentration bound derived from Hoeffding's Inequality and ϵ-cover are useless for problems with high dimensionality space such as Kernel Method.

We, therefore need a stronger and more general bound. We therefore study the McDiarmid Inequality, which is a more general result than the Hoeffding's Inequality.

2. The McDiarmid Inequality

The McDiarmid Inequality provides a more general bound than the Hoeffding's Inequality in the sense that it is independent of the dimensionality of any data point while computing the concentration bound, and thus can be used, even for high dimensional feature spaces.

Before we move on to understand the inequality, we must define what is c-stable function.

Definition 7.1 (c-stable function). Let $X_1, X_2 \ldots X_m \in \mathcal{X}^m$ be a set of $m \geq 1$ independent random variables and a function $f : \mathcal{X}^m \to \mathbb{R}$, then the function f is *c-stable* if $\exists\, c \in \mathbb{R}^+$ such that

$$| f(x_1 \ldots x_i \ldots x_m) - f(x_1 \ldots x_i' \ldots x_m) | \leq c \tag{1}$$

for all $i \in [\, 1, m\,]$ and any points $x_1, x_2 \ldots x_m, x_i' \in \mathcal{X}$.

We can consider a simple example for a function that is c-stable.

Example 7.1. Suppose we have a loss function $\mathcal{X}$ such that $\mathcal{X} \in [\, a, b\,]$ a.s. and a function $f : \mathcal{X}^m \to \mathbb{R}$ such that

$$f(x_1, x_2 \ldots x_n) \quad = \quad \frac{\sum_{i=1}^m x_i}{m}$$

then, we can argue that

$$
\begin{aligned}
| f(x_1 \ldots x_i \ldots x_m) - f(x_1 \ldots x_i' \ldots x_m) | \quad &= \quad \left| \frac{1}{m} \left[\sum_{j=1}^{m} x_j - \sum_{\substack{j=1 \\ j \neq i}}^{m} x_j - x_i' \right] \right| \\
&= \quad \frac{1}{m} \left| \left[\sum_{j=1}^{m} x_j - \sum_{j=1}^{m} x_j + x_i - x_i' \right] \right| \\
&= \quad \frac{1}{m} | x_i - x_i' | \\
&\leq \quad \frac{b-a}{m}
\end{aligned}
$$

Therefore, f is $\frac{b-a}{m}$ stable.

We can now state the McDiarmid Inequality, which is defined for only c-stable functions.

Theorem 7.1 (McDiarmid Inequality). Let $X_1, X_2 \ldots X_m \in \mathcal{X}^m$ be a set of $m \geq 1$ independent random variables and a function $f : \mathcal{X}^m \to \mathbb{R}$, such that the function f is c-stable. Let $f(S)$ denote $f(X_1, X_2 \ldots X_m)$, then the McDiarmid Inequality states that the for all $\epsilon > 0$,

$$
\mathbb{P}\left[\, | f(S) - \mathbb{E}\left[f(S) \right] | > \epsilon \,\right] \leq 2 \exp\left(\frac{-2\epsilon^2}{mc^2} \right) \tag{2}
$$

Proof. Let us first prove the inequality only for a samples with only a single variable, *i.e.* $m = 1$.

Suppose for a feature space $\mathcal{X}$, we have a random variable $X \in \mathcal{X}$ and a function $f : \mathcal{X} \to \mathbb{R}$ such that f is c-stable. Then, for all $x \in \mathcal{X}$, looking at only one side of the inequality (*i.e.* $f(X) - \mathbb{E}\left[f(X) \right] > \epsilon$), we can write

$$
\begin{aligned}
\mathbb{P}\left[\, f(X) - \mathbb{E}\left[f(X) \right] > \epsilon \,\right] \quad &= \quad \mathbb{P}\left[\, \exp\left(s \left(f(X) - \mathbb{E}\left[f(X) \right] \right) \right) > \exp\left(s\epsilon \right) \,\right] && [s > 0] \\
&\leq \quad \frac{\mathbb{E}\left[\, \exp\left(s \left(f(X) - \mathbb{E}\left[f(X) \right] \right) \right) \,\right]}{\exp\left(s\epsilon \right)} && [\text{Using Markov's Inequality}]
\end{aligned}
$$

Note that $\mathbb{E}\left[f(X) \right]$ is always a constant with respect to X. Therefore, we can say

$$
\begin{aligned}
\left(\max_X f(X) - \mathbb{E}\left[f(X) \right] \right) - \left(\min_X f(X) - \mathbb{E}\left[f(X) \right] \right) \quad &= \quad \left(\max_X f(X) \right) - \left(\min_X f(X) \right) \\
&= \quad \max_{X, X'} \left(f(X) - f(X') \right) \quad \leq \quad c
\end{aligned}
$$

Since f is c-stable, we can say that the difference between the maximum and the minimum values of the quantity $f(X) - \mathbb{E}\left[f(X) \right]$ is at most equal to c. Also, since $\mathbb{E}\left[f(X) - \mathbb{E}\left[f(X) \right] \right] = \mathbb{E}\left[f(X) \right] - \mathbb{E}\left[f(X) \right] = 0$.

Therefore, we can use Hoeffding's Lemma, discussed in Lecture XYZ, on the transformed random variable $f(X) - \mathbb{E}\left[f(X) \right]$. Hence, we get

$$
\mathbb{E}\left[\, \exp\left(s(f(X) - \mathbb{E}\left[f(X) \right]) \right) \,\right] \quad \leq \quad \exp\left(\frac{s^2 c^2}{8} \right)
$$

Putting this in the original inequality, we get

$$
\mathbb{P}\left[\, f(X) - \mathbb{E}\left[f(X) \right] > \epsilon \,\right] \quad \leq \quad \exp\left(\frac{s^2 c^2}{8} - s\epsilon \right)
$$

Since we want the closest bound, we minimize the RHS with respect to s. Note that this will be smallest for $s = 4\epsilon/c^2$, therefore

$$\mathbb{P}\left[\,f(X) - \mathbb{E}\left[\,f(X)\,\right] > \epsilon\,\right] \quad \leq \quad \exp\left(-\frac{2\epsilon^2}{c^2}\right)$$

We can follow the same set of steps to prove that $\mathbb{P}\left[\,f(X) - \mathbb{E}\left[\,X\,\right] < -\epsilon\,\right] \leq \exp\left(-2\epsilon^2/c^2\right)$. Combining these two inequalities, we get

$$\mathbb{P}\left[\,|\,f(X) - \mathbb{E}\left[\,f(X)\,\right]\,| > \epsilon\,\right] \quad \leq \quad 2\exp\left(-\frac{2\epsilon^2}{c^2}\right)$$

This is exactly the term of the McDiarmid Inequality when $m = 1$. Therefore, we can claim that the McDiarmid Inequality holds when $m = 1$. $\qquad \square$

In order to extend the proof to larger training samples, we need what is known as the Martingales. We shall look at this in the next section.

3. Martingales

In probability theory, a martingale is a sequence of random variables (i.e., a stochastic process) for which, at a particular time in the realized sequence, the expectation of the next value in the sequence is equal to the present observed value even given knowledge of all prior observed values. [1]

We define this more formally below.

Definition 7.2 (Martingale). A sequence of random variables $\{X_i\}_{i=1,2...}$ is said to be a *martingale*, if for all $i = 1, 2 \ldots$, the expectation $\mathbb{E}\left[\,|\,X_i\,|\,\right]$ exists and the conditional expectation

$$\mathbb{E}\left[\,X_{i+1} - X_i\,\middle|\,X_1, X_2 \ldots X_i\,\right] \quad = \quad 0\,.$$

In order to better understand this, consider the following example of a martingale sequence.

Example 7.2 (Random Walk). Suppose a drunkard started his walk from a fixed, known position (say the origin) in 1-D. At any second, he may choose one direction, *i.e.* left or right at random, and take one step and one step only. If we define the position of the drunkard (a random variable) at any second i as Z_i (let Z_1 be the initial position), then we can say that the sequence $\{Z_i\}_{i=1,2...}$ is a martingale sequence.

It is easy to see why indeed this is a martingale sequence. We can write the expectation

$$\begin{aligned}
\mathbb{E}\left[\,Z_{i+1} - Z_i\,\middle|\,Z_1, Z_2 \ldots Z_i\,\right] \quad &= \quad \mathbb{E}\left[\,Z_{i+1}\,\middle|\,Z_i\,\right] - Z_n \\
&= \quad \frac{1}{2}(Z_i + 1) + \frac{1}{2}Z_i - 1 - Z_i \\
&= \quad 0
\end{aligned}$$

Since the condition for this sequence to be a martingale holds true and therefore, the random walk sequence is a martingale sequence.

Definition 7.3 (Martingale Difference). A sequence of random variables $\{Z_i\}_{i=1,2...}$ is said to be a *martingale difference* with respect to another sequence of random variables $\{X_i\}_{i=1,2...}$ if for all $i = 1, 2 \ldots$, Z_i is a function $X_1, X_2 \ldots X_i$, $\mathbb{E}\left[\,|\,Z_i\,|\,\right]$ exists, and

$$\mathbb{E}\left[\,Z_{i+1} - Z_{i+1}\,\middle|\,X_1, X_2 \ldots X_i\,\right] \quad = \quad 0$$

We will derive a concentration bound for a sequence that is a martingale difference with respect to another sequence. Using this, we can prove the McDiarmid Inequality.

Theorem 7.2. Let $\{Z_i\}_{i=1,2\ldots}$ be a martingale difference with respect to another sequence $\{X_i\}_{i=1,2\ldots}$. Suppose $\exists\, c \in \mathbb{R}^+$ such that for all $i = 1, 2\ldots$

$$a \quad \leq \quad Z_{i+1} - Z_i \quad \leq \quad b$$

then, for all $\epsilon > 0$,

$$\mathbb{P}\left[\,|\, Z_{n+1} - Z_1 \,| > \epsilon\,\right] \quad \leq \quad 2 \exp\left(-\frac{2\epsilon}{n(b-a)^2}\right) \tag{3}$$

Proof. We borrow the proof technique from the proofs of the Multiplicative Chernoff's Bound and Hoeffding's Bound.

$$
\begin{aligned}
\mathbb{P}\left[\,(Z_n - Z_1) > \epsilon\,\right] &= \mathbb{P}\left[\,s\,(Z_n - Z_1) > s\epsilon\,\right] \\
&= \mathbb{P}\left[\,\exp\left(s\,(Z_{n+1} - Z_1)\right) > \exp\left(s\epsilon\right)\,\right] \\
&\leq \frac{\mathbb{E}\left[\,\exp\left(s\,(Z_{n+1} - Z_1)\right)\,\right]}{\exp\left(s\epsilon\right)}
\end{aligned}
$$

Let us represent $\{X_j\}_{j=1,2\ldots i}$ as $\overline{X}_i$.

Since for all $i = 1, 2\ldots$, the random variable Z_i is a function of $\overline{X}_i$ (from the definition of martingales), we can therefore write the expectation term as below

$$
\begin{aligned}
\mathbb{E}\left[\,\exp\left(s\,(Z_{n+1} - Z_1)\right)\,\right] &= \mathop{\mathbb{E}}_{\overline{X}_{n+1}}\left[\exp\left(s\left(\sum_{i=1}^{n} Z_{i+1} - Z_i\right)\right)\right] \\
&\leq \mathop{\mathbb{E}}_{\overline{X}_{n+1}}\left[\prod_{i=1}^{n}\exp\left(s\,(Z_{i+1} - Z_i)\right)\right] \\
&= \mathop{\mathbb{E}}_{\overline{X}_{n}}\left[\mathop{\mathbb{E}}_{X_{n+1}}\left[\prod_{i=1}^{n}\exp\left(t\,(Z_{i+1} - Z_i)\right) \,|\, \overline{X}_n\right]\right]
\end{aligned}
$$

Since for all $i = 1, 2\ldots n$ we can say Z_j is a function of $\overline{X}_n$, therefore, the product terms for $i = 1\ldots n-1$ will be constant with respect to the expectation, and hence will come out of the inner expectation.

$$\mathbb{E}\left[\,\exp\left(s\,(Z_{n+1} - Z_1)\right)\,\right] \quad \leq \quad \mathop{\mathbb{E}}_{\overline{X}_n}\left[\prod_{i=1}^{n-1}\exp\left(s\,(Z_{i+1} - Z_i)\right)\mathop{\mathbb{E}}_{X_{n+1}}\left[\,\exp\left(s\,(Z_{n+1} - Z_n)\right) \,|\, \overline{X}_n\,\right]\right]$$

We can apply Hoeffding's Lemma on $(Z_{n+1} - Z_n)$ as given $\overline{X}_n$, we know $(Z_n - Z_1) \in [\,a, b\,]$ and $\mathbb{E}\left[\,Z_{n+1} - Z_n \,|\, \overline{X}_n\,\right] = 0$. Therefore

$$\mathbb{E}\left[\,\exp\left(s\,(Z_{n+1} - Z_n)\right) \,|\, \overline{X}_n\,\right] \quad \leq \quad \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

Since, the expectation will always be non-negative, the inequality will hold, and hence, we can write

$$\mathbb{E}\left[\,\exp\left(s\,(Z_{n+1} - Z_n)\right) \,|\, \overline{X}_n\,\right] \quad \leq \quad \mathop{\mathbb{E}}_{\overline{X}_n}\left[\prod_{i=1}^{n-1}\exp\left(s\,(Z_{i+1} - Z_i)\right) \,|\, \overline{X}_n\right]\exp\left(\frac{s^2(b-a)^2}{8}\right)$$

Iterating the same steps, we can therefore reduce the inequality to as written below.

$$\mathbb{E}\left[\,\exp\left(s\,(Z_n - Z_1)\right)\,\right] \quad \leq \quad \exp\left(\frac{ns^2(b-a)^2}{8}\right)$$

4

Formulating the detailed steps for above is left as an exercise. Putting this in the original equation, we get

$$\mathbb{P}\left[(Z_n - Z_1) > \epsilon\right] \quad \leq \quad \exp\left(\frac{ns^2(b-a)^2}{8} - s\epsilon\right)$$

Minimizing the upper bound with respect to s, we get $s = 4\epsilon/(b-a)^2$, therefore

$$\mathbb{P}\left[(Z_n - Z_1) > \epsilon\right] \quad \leq \quad \exp\left(-\frac{2\epsilon^2}{n(b-a)^2}\right)$$

We can follow the steps to prove that

$$\mathbb{P}\left[Z_{n+1} - Z_1 < \epsilon\right] \quad \leq \quad \exp\left(-\frac{2\epsilon^2}{n(b-a)^2}\right)$$

Therefore, combining the two inequalities, we get

$$\mathbb{P}\left[\,|\,Z_{n+1} - Z_1\,| > \epsilon\right] \quad \leq \quad 2\exp\left(-\frac{2\epsilon^2}{n(b-a)^2}\right)$$

$\square$

Note. The above theorem is a direct corollary of Azuma's Inequality (see [2])

Using this theorem, we can finally prove the McDiarmid Inequality, as we will see in the next section

4.   Proof and Corollaries of the McDiarmid Inequality

We can now prove the McDiarmid Inequality using the concept of martingales.

Note. We follow the same notation from the previous section, *i.e.* $\overline{X}_n$ represents $X_1, X_2 \ldots X_n$

Following from the statement of the McDiarmid Inequality, let $X_1, X_2 \ldots X_n \in \mathcal{X}^n$ be a set of independent random variables and $f : \mathcal{X}^n \to \mathbb{R}$ be a c-stable function. Assume another sequence of random variables, $\overline{Z}_{n+1}$ such that for all $i = 0, 1 \ldots n-1$

$$Z_i \quad = \quad \underset{X_{i+1} \ldots X_n}{\mathbb{E}}\left[f(X_1, X_2 \ldots X_n)\,\middle|\,\overline{X}_i\right]$$

and $Z_n = f(X_1, X_2 \ldots X_n)$

We therefore want to show

$$\mathbb{E}\left[\,|\,Z_n - Z_0\,| > \epsilon\right] \quad \leq \quad 2\exp\left(-\frac{2\epsilon^2}{nc^2}\right)$$

Before showing this, let us consider the following claim.

Claim 7.3. The sequence $\overline{Z}_n$ is a martingale difference with respect to $\overline{X}_n$.

Proof.   In order to prove that $\overline{Z}_n$ is a martingale difference with respect to $\overline{X}_n$, we need to show

$$\mathbb{E}\left[Z_{i+1} - Z_i\,\middle|\,\overline{X}_i\right] = 0$$

For simplicity, we represent $X_{i+1} \ldots X_n$ as $\overline{X}_{-i}$

$$\mathbb{E}\left[\,Z_{i+1} - Z_i \,\middle|\, \overline{X}_i\,\right] \;=\; \mathbb{E}\left[\,Z_{i+1}\,\middle|\,\overline{X}_i\,\right] - \mathbb{E}\left[\,Z_i\,\middle|\,\overline{X}_i\,\right]$$

$$=\; \mathop{\mathbb{E}}_{\overline{X}_{-i}}\left[\,Z_{i+1}\,\middle|\,\overline{X}_i\,\right] - \mathop{\mathbb{E}}_{\overline{X}_{-i}}\left[\,Z_i\,\middle|\,\overline{X}_i\,\right]$$

$$=\; \mathop{\mathbb{E}}_{\overline{X}_{-i}}\left[\,\mathop{\mathbb{E}}_{\overline{X}_{-(i+1)}}\left[\,f(X_1\dots X_n)\,\middle|\,\overline{X}_{i+1}\,\right]\,\middle|\,\overline{X}_i\,\right] - Z_i$$

From the properties of conditional expectation,

$$\mathbb{E}\left[\,A\,\middle|\,B\,\right] \;=\; \mathbb{E}\left[\,\mathbb{E}\left[\,A\,\middle|\,B,C\,\right]\,\middle|\,B\,\right]$$

Therefore,

$$\mathop{\mathbb{E}}_{\overline{X}_{-i}}\left[\,\mathop{\mathbb{E}}_{\overline{X}_{-(i+1)}}\left[\,f(X_1\dots X_n)\,\middle|\,\overline{X}_{i+1}\,\right]\,\middle|\,\overline{X}_i\,\right] \;=\; \mathop{\mathbb{E}}_{\overline{X}_{-i}}\left[\,f(X_1\dots X_n)\,\middle|\,\overline{X}_i\,\right]$$

$$=\; Z_i$$

$$\implies \mathbb{E}\left[\,Z_{i+1} - Z_i\,\middle|\,\overline{X}_i\,\right] \;=\; 0$$

Hence, our claim is true. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Remark. We can also include X_0 and Z_0 in the sequence where X_0 can be any random variable. Since Z_1 is independent of X_0, $\mathbb{E}\left[\,Z_1 - Z_0\,\middle|\,X_0\,\right] = \mathbb{E}\left[\,Z_1 - Z_0\,\right] = 0$. Hence, the sequence $Z_0, Z_1\dots Z_n$ is a martingale difference with respect to the sequence $X_0, X_1\dots X_n$.

Exercise 7.1. Prove that since f is c-stable, for all $i = 0, 1\dots n-1$, $|\,Z_{i+1} - Z_i\,| \le c$.

Since we have shown that the sequence $Z_0\dots Z_n$ is a martingale difference with respect to the sequence $X_0\dots X_n$, we can use Theorem 7.2. Also from Exercise 7.1, we know that $|\,Z_{i+1} - Z_i\,| \le c$, we can apply Theorem 7.2. From that, we get

$$\mathbb{P}\left[\,|\,Z_n - Z_0\,| > \epsilon\,\right] \;\le\; 2\exp\left(-\frac{\epsilon^2}{2nc^2}\right)$$

However, note that this is not the same as the expression given for the McDiarmid Inequality. We can find a tighter bound by realizing the difference between the maximum and the minimum values of $Z_{i+1} - Z_i$ given $\overline{X}_i$.

Let $W_i = \sup\{Z_{i+1} - Z_i\}$ given $\overline{X}_i$ and $U_i = \inf\{Z_{i+1} - Z_i\}$

Since we are given $\overline{X}_i$, therefore Z_i is a constant (as Z_i is a function of $\overline{X}_i$). Also, since we are computing expectation over $\overline{X}_{-(i+1)}$ in Z_{i+1}, the only variable in question is X_{i+1}. Therefore, we can rewrite W_i and U_i as

$$W_i \;=\; \sup_{x}\left\{\mathop{\mathbb{E}}_{\overline{X}_{-(i+1)}}\left[\,f(S)\,\middle|\,X_1, X_2\dots X_i, x\,\right]\right\} - Z_i$$

$$U_i \;=\; \inf_{x}\left\{\mathop{\mathbb{E}}_{\overline{X}_{-(i+1)}}\left[\,f(S)\,\middle|\,X_1, X_2\dots X_i, x\,\right]\right\} - Z_i$$

Hence, the difference can be written as

$$\implies W_i - U_i \;=\; \sup_{x}\left\{\mathop{\mathbb{E}}_{\overline{X}_{-(i+1)}}\left[\,f(S)\,\middle|\,X_1, X_2\dots X_i, x\,\right]\right\} - \inf_{x}\left\{\mathop{\mathbb{E}}_{\overline{X}_{-(i+1)}}\left[\,f(S)\,\middle|\,X_1, X_2\dots X_i, x\,\right]\right\}$$

$$=\; \sup_{x,x'}\left\{\mathop{\mathbb{E}}_{\overline{X}_{-(i+1)}}\left[\,f(S)\,\middle|\,X_1, X_2\dots X_i, x\,\right] - \mathop{\mathbb{E}}_{\overline{X}_{-(i+1)}}\left[\,f(S)\,\middle|\,X_1, X_2\dots X_i, x'\,\right]\right\}$$

This, similar to Exercise 7.1 can be proved to be smaller than c. Also,

$$U_i \quad \leq \quad Z_{i+1} - Z_i \quad \leq \quad W_i$$

Since U_i is a constant given $\overline{X}_i$ as it does not depend on $\overline{X}_{-i}$, and $W_i - U_i \leq c$, we have

$$U_i \quad \leq \quad Z_{i+1} - Z_i \quad \leq \quad U_i + c$$

Applying Theorem 7.2, we get the actual term for the McDiarmid Inequality, *i.e.*

$$\mathbb{P}\left[\,|\,Z_n - Z_0\,|\,\right] \quad \leq \quad \exp\left(-\frac{2n\epsilon^2}{c^2}\right)$$

$$\implies \mathbb{P}\left[\,|\,f(X_1, X_2 \ldots X_n) - \mathbb{E}\left[\,f(X_1, X_2 \ldots X_n)\,\right]\,|\,\right] \quad \leq \quad \exp\left(-\frac{2n\epsilon^2}{c^2}\right)$$

This completes the proof of the McDiarmid Inequality. Below, we see some interesting applications / corollaries of the McDiarmid Inequality.

Remark. We can use the McDiarmid Inequality for functions which are c-stable with high probability, but not almost surely.

Corollary 7.1.1 (Hoeffding's Inequality). Consider the function discussed in Example 7.1., *i.e.* $f(X_1 \ldots X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$. Since, f is $\frac{b-a}{n}$–stable, we can apply the the McDiarmid Inequality to f, using which we directly get the Hoeffding's Inequality. The details of the intermediate steps are left as an exercise.

Corollary 7.1.2 (Uniform Convergence). If for some loss function $l : \mathcal{X} \times \mathcal{Y}$, we have $\forall\,(x,y) \in \mathcal{X} \times \mathcal{Y}$ and $\forall\,f$, $l(f(x), y) \in [\,a, b\,]$, then the uniform condition holds as follows

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}} \left\{\mathrm{er}_S^l\,[\,f\,] - \mathrm{er}_D^l\,[\,f\,]\right\} > \epsilon\right] \quad \leq \quad \exp\left(-\frac{2n\epsilon}{(b-a)^2}\right)$$

Proof. Suppose we have a loss function $l : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, such that for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and $\forall\,f \in \mathcal{F}$, we have

$$a \quad \leq \quad l(f(x), y) \quad \leq \quad b$$

Suppose we have a training sample $S = \{(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)\}$. Let

$$g(S) \quad = \quad \sup_{f \in \mathcal{F}} \left\{\mathrm{er}_S^l\,[\,f\,] - \mathrm{er}_D^l\,[\,f\,]\right\}$$

We want to show that g is c-stable, which would allow us to use the McDiarmid Inequality for this function. Note that the expectation value of g is zero.

Define $g_f(S) = \mathrm{er}_S^l\,[\,f\,] - \mathrm{er}_D^l\,[\,f\,]$ and $S_i = S \setminus \{(x_i, y_i)\} \bigcup \{(x_i', y_i)\}$

Remark. $g(S)$, information theoretical wise, is well defined, even though we cannot calculate the value of $g(S)$ algorithmically.

Since $l(f(x), y) \in [\,a, b\,]$, we have for all $i \in [\,n\,]$

$$g_f(S) - g_f(S_i) \quad = \quad \frac{1}{n}\left(\sum_{j=1}^{n} l(f(x_j), y_j) - \left(\sum_{\substack{j=1 \\ j \neq i}}^{n} l(f(x_j), y_j) + l(f(x_i'), y_i)\right)\right)$$

$$= \quad \frac{1}{n}\left(l(f(x_i), y_i) - l(f(x_i'), y_i)\right)$$

7

Since $l(f(x), y) \in [\,a, b\,]$

$$g_f(S) - g_f(S_i) \quad \leq \quad \frac{b-a}{n} \tag{4}$$

Therefore, we can write, for all $i \in [\,n\,]$

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

Suppose $g(S)$ is maximum for f_1 and $g(S_i)$ is maximum for f_2 (without loss of generality)

$$
\begin{aligned}
g(S) - g(S_i) \quad &= \quad g_{f_1}(S) - g_{f_2}(S_i) \\
&\leq \quad g_{f_1}(S) - g_{f_1}(S_i) + \frac{b-a}{n} \qquad\qquad \text{[Using Equation 4]} \\
&= \quad \frac{b-a}{n}
\end{aligned}
$$

Therefore, g is $\frac{b-a}{n}$-stable. This suggests that we can use the McDiarmid Inequality for g. Hence, we get

$$\mathbb{P}\left[g(S) - \mathbb{E}_S[\,g(S)\,] > \epsilon \right] \quad \leq \quad \exp\left(-\frac{2n\epsilon^2}{(b-a)^2} \right)$$

$$\implies \mathbb{P}\left[\sup_{f \in \mathcal{F}} \{ \mathrm{er}^l_S[\,f\,] - \mathrm{er}^l_D[\,f\,] \} > \epsilon \right] \quad \leq \quad \exp\left(-\frac{2n\epsilon^2}{(b-a)^2} \right)$$

We can also prove the negative tail for this, using the same set of steps. Therefore, we get the bound for Uniform Convergence using the McDiarmid Inequality. $\qquad\square$

We can therefore see that the McDiarmid Inequality is really useful for finding bounds for c-stable functions, even for infinite hypothesis classes, since the bound is independent of the size of the hypothesis class.

In the next lecture, we will revise the bound for the regression case using the McDiarmid Inequality and see that for certain hypothesis classes, it is possible to get a bound on the number of samples required for a model ϵ-close to the real data distribution.

References

[1] Wikipedia *Martingale, Probability Theory*.
   https://en.wikipedia.org/wiki/Martingale_(probability_theory)

[2] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar *Foundations of Machine Learning, 2012*