1 McDiarmid’s inequality

First, we start with some necessary definitions.

**Definition 5.1.** Let \( f \) be a function of two random variables \( x \) and \( y \) with support on \( \mathcal{X} \) and \( \mathcal{Y} \) respectively. The conditional expectation of \( f(x, y) \) given \( y \) is defined as:

\[
E_x[f(x, y)|y] = \int_{x \in \mathcal{X}} f(x, y)p(x|y)dx
\]

The conditional expectation has some natural properties such as the following. The law of total expectation states that:

\[
E_y[E_x[f(x, y)|y]] = E_{xy}[f(x, y)]
\]

A generalization is the law of iterated expectation, which states that:

\[
E_y[E_x[f(x, y, z)|y, z]|z] = E_{xyz}[f(x, y, z)|z]
\]

(The proof of the above might be left for homework very soon.)

In what follows, we present and prove the main result.

**Theorem 5.1** (McDiarmid’s inequality). Let \( x_1 \ldots x_n \) be independent random variables with support on \( \mathcal{X} \). Let \( f: \mathcal{X}^n \to \mathbb{R} \) be a function that satisfies the following condition:

\[
(\forall i, \forall x_1 \ldots x_n, x'_i) \ |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x'_i, \ldots, x_n)| \leq c_i \quad (1)
\]

Fix \( \varepsilon > 0 \), we have that:

\[
P[f(x_1, \ldots, x_n) - E[f(x_1, \ldots, x_n)] \geq \varepsilon] \leq e^{-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}}
\]

**Proof.** Let \( x'_i \equiv x_1 \ldots x_j \) and \( x \equiv x_1^n \). Define the random variable:

\[
z_i \equiv E_{x_{i+1}^n}[f(x)|x_i^j]
\]
We will make several observations regarding the above random variable. First, note that:

\[ z_n = E_x^n [f(x)|x^n_1] = f(x) \]
\[ z_0 = E_x^0 [f(x)|x^0_0] = E_x[f(x)] \]
\[ f(x) - E_x[f(x)] = z_n - z_0 = \sum_{i=1}^{n} (z_i - z_{i-1}) \quad (2) \]

Second, by the law of iterated expectations, for all \( m \):

\[ E_{x_1^{m-1}}[E_{x_m}[e^{t\sum_{i=1}^{m} (z_i - z_{i-1})}|x_1^{m-1}]] = E_{x_1^{m-1}}[e^{t\sum_{i=1}^{m} (z_i - z_{i-1})}] \quad (3) \]

Third, consider the random variable \( z_i - z_{i-1}|x_1^{i-1} \). By the law of iterated expectations, for all \( i \):

\[ E_{x_i}[z_i - z_{i-1}|x_1^{i-1}] = E_{x_i}[z_i|x_1^{i-1}] - E_{x_i}[z_{i-1}|x_1^{i-1}] \]
\[ = E_{x_i}[\mathbb{E}_{x_{i+1}^n} f(x)|x_1^{i-1}] - E_{x_i}[\mathbb{E}_{x_{i}^n} f(x)|x_1^{i-1}] \]
\[ = E_{x_i}[\mathbb{E}_{x_{i+1}^n} f(x)|x_1^{i-1}, x_i] - E_{x_i}[\mathbb{E}_{x_{i}^n} f(x)|x_1^{i-1}] \]
\[ = E_{x_i}[\mathbb{E}_{x_{i}^n} f(x)|x_1^{i-1}] - E_{x_i}[\mathbb{E}_{x_{i}^n} f(x)|x_1^{i-1}] \]
\[ = 0 \quad (4) \]

Furthermore:\(^1\)

\[ z_i - z_{i-1}|x_1^{i-1} = E_{x_i}[f(x)|x_1^{i}] - E_{x_i}[f(x)|x_1^{i-1}] \]
\[ = E_{x_i}[\mathbb{E}_{x_{i+1}^n} f(x)|x_1^{i-1}, x_i] - E_{x_i}[\mathbb{E}_{x_{i}^n} f(x)|x_1^{i-1}] \]
\[ \leq \sup_{x_i} \left( E_{x_i}[\mathbb{E}_{x_{i+1}^n} f(x_1 \ldots x_i \ldots x_n)|x_1^{i-1}, x_i] - E_{x_i}[\mathbb{E}_{x_{i}^n} f(x)|x_1^{i-1}] \right) \]
\[ \equiv b_i \]

and similarly:\(^2\)

\[ z_i - z_{i-1}|x_1^{i-1} = \mathbb{E}_{x_i}[f(x)|x_1^{i}] - \mathbb{E}_{x_i}[f(x)|x_1^{i-1}] \]
\[ = \mathbb{E}_{x_i}[\mathbb{E}_{x_{i+1}^n} f(x)|x_1^{i-1}, x_i] - \mathbb{E}_{x_i}[\mathbb{E}_{x_{i}^n} f(x)|x_1^{i-1}] \]
\[ \geq \inf_{x_i} \left( \mathbb{E}_{x_i}[\mathbb{E}_{x_{i+1}^n} f(x_1 \ldots x_i \ldots x_n)|x_1^{i-1}, x_i] - \mathbb{E}_{x_i}[\mathbb{E}_{x_{i}^n} f(x)|x_1^{i-1}] \right) \]
\[ \equiv a_i \]

Note that, by eq.(1) we have \( b_i - a_i \leq c_i \). Therefore by eq.(4) and Hoeffding’s lemma (by change of variables from \([a_i, b_i]\) to \([0, 1]\) as we did in Homework 1 Question 3), we have for all \( t \in \mathbb{R} \):

\[ E_{x_i}[e^{t(z_i - z_{i-1})}|x_1^{i-1}] \leq e^{\frac{1}{2}t^2(b_i - a_i)^2} \leq e^{\frac{1}{2}t^2 c_i^2} \quad (5) \]

\(^1\)If you are not familiar with supremum (sup), you can intuitively use the maximum (max).
\(^2\)If you are not familiar with infimum (inf), you can intuitively use the minimum (min).
By using these observations, we use the techniques that we have used in previous lectures. Let \( t > 0 \), by Markov’s inequality (Theorem 1.1), eq.(2), eq.(3) and eq.(5), we have:

\[
\mathbb{P}_x[f(x) - \mathbb{E}_x[f(x)] \geq \varepsilon] = \mathbb{P}_x[e^{t(f(x) - \mathbb{E}_x[f(x)])} \geq e^{t\varepsilon}]
\leq e^{-t\varepsilon} \mathbb{E}_x[e^{t(f(x) - \mathbb{E}_x[f(x)])}]
= e^{-t\varepsilon} \mathbb{E}_x[e^{t\sum_{i=1}^n (x_i-z_i-1)}]
= e^{-t\varepsilon} \mathbb{E}_{x_i=1}^n \mathbb{E}_{x_n}[e^{t\sum_{i=1}^{n-1} (z_i-z_{i-1})} | x_1^{n-1}]]
= e^{-t\varepsilon} \mathbb{E}_{x_i=1}^n \mathbb{E}_{x_n}[e^{t(x_n-z_{n-1})} | x_1^{n-1}]]
\leq e^{-t\varepsilon + \frac{1}{2}t^2 c_i^2} \mathbb{E}_{x_i=1}^n \mathbb{E}_{x_n}[e^{t(x_n-z_{n-1})}]
\vdots
\leq e^{-t\varepsilon + \frac{1}{2}t^2 (\sum_{i=1}^n c_i^2)}
\]

In order to minimize the quadratic function \( f(t) = -t\varepsilon + \frac{1}{8}t^2 (\sum_{i=1}^n c_i^2) \), we make the derivative equal to zero and solve for \( t \). That is:

\[
0 = \partial f(t)/\partial t
= -\varepsilon + \frac{1}{4}t \left( \sum_{i=1}^n c_i^2 \right)
\]

Thus, \( t = 4\varepsilon / (\sum_{i=1}^n c_i^2) \). Plugging this back in the above, we prove our claim. 

The above theorem is remarkably general. Some of the results from Lecture 2 can be proved by using Theorem 5.1.

**Corollary 5.1 (Hoeffding’s inequality).** Assume that \( x_1 \ldots x_n \) are \( n \) independent random variables, where each \( x_i \in [a_i, b_i] \). Fix \( \varepsilon > 0 \). We have that:

\[
\mathbb{P} \left[ \frac{1}{n} \sum_{i=1}^n x_i - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n x_i \right] \geq \varepsilon \right] \leq e^{\frac{-2n^2 \varepsilon^2}{\sum_{i=1}^n (b_i-a_i)^2}}
\]

**Proof.** Let \( f(x_1 \ldots x_n) = \frac{1}{n} \sum_{i=1}^n x_i \). As in eq.(1), we have:

\[
(\forall i, \forall x_1 \ldots x_n, x'_i) \ |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x'_i, \ldots, x_n)|
= \left| \frac{1}{n} \left( \sum_{j=1}^{i-1} x_j + x_i + \sum_{j=i+1}^n x_j \right) - \frac{1}{n} \left( \sum_{j=1}^{i-1} x_j + x'_i + \sum_{j=i+1}^n x_j \right) \right|
= \left| \frac{1}{n} (x_i - x'_i) \right|
\leq \frac{b_i - a_i}{n}
\]

By invoking Theorem 5.1 with \( c_i = (b_i - a_i)/n \), we prove our claim. 

\[ \square \]
2 Sub-Gaussian Random Variables

**Definition 5.2.** A random variable \( x \) with support on \( \mathbb{R} \) and mean \( \mu = \mathbb{E}[x] \) is called sub-Gaussian with parameter \( \sigma^2 \), if its moment generating function is bounded as follows:

\[
\mathbb{E}[e^{t(x-\mu)}] \leq e^{\frac{\sigma^2 t^2}{2}}
\]

for all \( t \in \mathbb{R} \).

Gaussian random variables with variance \( \sigma^2 \) are sub-Gaussian with parameter \( \sigma^2 \) (For instance, see proof of Corollary 1.2). If \( x \sim \mathcal{N}(\mu, \sigma^2) \), we have:

\[
\mathbb{E}[e^{t(x-\mu)}] = \int_{-\infty}^{+\infty} e^{t(x-\mu)} \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} dx = e^{\frac{\sigma^2 t^2}{2}}
\]

Random variables with support on \([0, 1]\) are sub-Gaussian with parameter \( (\frac{1}{2})^2 \).

As stated in Hoeffding’s Lemma 2.1:

\[
\mathbb{E}[e^{t(x-\mu)}] \leq e^{\frac{1}{8} t^2}
\]

Random variables with support on \([a, b]\) are sub-Gaussian with parameter \( (\frac{b-a}{2})^2 \) (See solution of Homework 1 Question 3).

Next, we present some properties of sub-Gaussian random variables that are easy to prove.

**Claim 5.1.** Let \( a \) be a constant and let \( x \) be a sub-Gaussian random variable with mean \( \mu \) and parameter \( \sigma^2 \). The random variable \( ax \) is sub-Gaussian with mean \( a\mu \) and parameter \( (a\sigma)^2 \).

**Proof.** First, the mean is given by \( \mathbb{E}[ax] = a\mathbb{E}[x] \). Then, for all \( t \in \mathbb{R} \), we have:

\[
\mathbb{E}[e^{t(ax-\mathbb{E}[ax])}] = \mathbb{E}[e^{ta(x-\mathbb{E}[x])}]
\]

\[
\leq e^{\frac{(a\sigma)^2 t^2}{2}}
\]

\( \square \)

**Claim 5.2.** If \( x \) and \( y \) are two independent sub-Gaussian random variables with mean \( \mu_x \) and \( \mu_y \) and parameter \( \sigma_x^2 \) and \( \sigma_y^2 \) respectively, then \( x + y \) is sub-Gaussian with mean \( \mu_x + \mu_y \) and parameter \( \sigma_x^2 + \sigma_y^2 \).

**Proof.** First, the mean is given by \( \mathbb{E}[x+y] = \mathbb{E}[x] + \mathbb{E}[y] \). Then, for all \( t \in \mathbb{R} \), we have:

\[
\mathbb{E}[e^{t(x+y-\mathbb{E}[x+y])}] = \mathbb{E}[e^{t(x-\mathbb{E}[x])}e^{t(y-\mathbb{E}[y])}]
\]

\[
= \mathbb{E}[e^{t(x-\mathbb{E}[x])}] \mathbb{E}[e^{t(y-\mathbb{E}[y])}]
\]

\[
\leq e^{\frac{\sigma_x^2 t^2}{2}} e^{\frac{\sigma_y^2 t^2}{2}}
\]

\[
= e^{\frac{(\sigma_x^2 + \sigma_y^2) t^2}{2}}
\]

\( \square \)
Claim 5.3. If \( x \) is a sub-Gaussian random variable with mean \( \mu \) and parameter \( \sigma^2 \), then:

\[
\begin{align*}
\mathbb{P}[x - \mu \geq \varepsilon] &\leq e^{\frac{-\varepsilon^2}{2\sigma^2}} \\
\mathbb{P}[\mu - x \geq \varepsilon] &\leq e^{\frac{-\varepsilon^2}{2\sigma^2}} \\
\mathbb{P}[|x - \mu| \geq \varepsilon] &\leq 2e^{\frac{-\varepsilon^2}{2\sigma^2}}
\end{align*}
\]

Proof. Pick some arbitrary \( t > 0 \). By Markov’s inequality (Theorem 1.1), we have:

\[
\begin{align*}
\mathbb{P}[x - \mu \geq \varepsilon] &= \mathbb{P}[e^{t(x-\mu)} \geq e^{t\varepsilon}] \\
&\leq e^{-t\varepsilon} \mathbb{E}[e^{t(x-\mu)}] \\
&\leq e^{-t\varepsilon} e^{\frac{\varepsilon^2}{2\sigma^2}} \\
&= e^{-\varepsilon + \frac{\varepsilon^2}{2\sigma^2}}
\end{align*}
\]

In order to minimize the quadratic function \( f(t) = -t\varepsilon + \frac{\varepsilon^2}{2\sigma^2} \), we make the derivative equal to zero and solve for \( t \). That is:

\[
0 = \partial f(t)/\partial t \\
= -\varepsilon + \sigma^2 t
\]

Thus, \( t = \varepsilon/\sigma^2 \). Plugging this back in the above, we prove our first claim.

Similarly, by Markov’s inequality (Theorem 1.1), we have:

\[
\begin{align*}
\mathbb{P}[\mu - x \geq \varepsilon] &= \mathbb{P}[e^{t(\mu-x)} \geq e^{t\varepsilon}] \\
&\leq e^{-t\varepsilon} \mathbb{E}[e^{t(\mu-x)}] \\
&\leq e^{-t\varepsilon} \mathbb{E}[e^{-t(x-\mu)}] \\
&\leq e^{-t\varepsilon} e^{\frac{\varepsilon^2}{2\sigma^2}} \\
&= e^{-\varepsilon + \frac{\varepsilon^2}{2\sigma^2}}
\end{align*}
\]

We continue as before and we prove the second claim. Finally, by the union bound, we prove the third claim. \( \square \)

Before going into more developed properties of sub-Gaussian random variables, we revise some general definitions.

**Theorem 5.2** (Layer cake representation). Let \( x \) be a non-negative random variable. We have that:

\[
(\forall r > 0) \ \mathbb{E}[x^r] = \int_0^{+\infty} r t^{r-1} \mathbb{P}[x \geq t] dt
\]
Proof. Note that \( x^r = r \int_0^x t^{r-1} dt \). Therefore:

\[
E[x^r] = \int_0^{+\infty} x^r p(x) dx \\
= \int_0^{+\infty} \left( r \int_0^x t^{r-1} dt \right) p(x) dx \\
= \int_0^{+\infty} \left( r \int_0^{+\infty} t^{r-1} 1[t \leq x] dt \right) p(x) dx \\
= \int_0^{+\infty} r t^{r-1} \left( \int_0^{+\infty} 1[t \leq x] p(x) dx \right) dt \\
= \int_0^{+\infty} r t^{r-1} P[x \geq t] dt
\]

\( \Box \)

Definition 5.3. The Gamma function is defined as \( \Gamma(r) = \int_0^{+\infty} u^{r-1} e^{-u} du \) for \( r > 0 \). For a positive integer \( r \) we have \( \Gamma(r) = (r-1)! \).

Next, we show that all moments of sub-Gaussians are bounded.

Theorem 5.3 (Bounded moments for sub-Gaussians). Let \( x \) be a sub-Gaussian random variable with mean \( \mu \) and parameter \( \sigma^2 \). We have:

\[
(\forall r > 0) \quad E[|x - \mu|^r] \leq r^{2r/2}\sigma^r \Gamma(r/2)
\]

Proof. Let \( u = \frac{\mu^2}{2\sigma^2} \) then:

\[
t = \sigma(2u)^{1/2} \\
t^{r-1} dt = \sigma^{r-1}(2u)^{(r-1)/2} \left( \frac{\sigma}{(2u)^{1/2}} du \right) \\
= \sigma^r(2u)^{r/2-1} du
\]

Note that by the layer cake representation (Theorem 5.2) and sub-Gaussianity:

\[
E[|x - \mu|^r] = \int_0^{+\infty} rt^{r-1} P[|x - \mu| \geq t] dt \\
\leq \int_0^{+\infty} rt^{r-1} 2e^{-\frac{t^2}{2\sigma^2}} dt \\
= 2r \int_0^{+\infty} t^{r-1} e^{-\frac{t^2}{2\sigma^2}} dt \\
= 2r \int_0^{+\infty} \sigma^r(2u)^{r/2-1} e^{-u} du \\
= r2^{r/2}\sigma^r \Gamma(r/2)
\]

\( \Box \)