5.1 Martingale-based methods

Last time, we studied tail bounds on the maximum of random variables as well as a quadratic form of random variables. Now we turn our attention to concentration inequalities of more general functions.

5.1.1 Bounded difference inequality

**Theorem 5.1** Let \( \{D_k, \mathcal{F}_k\}_{k=1}^\infty \) be a martingale difference sequence and suppose that \( \mathbb{E} \left[ e^{\lambda D_k} \mid \mathcal{F}_{k-1} \right] \leq e^{\lambda^2 \nu_k^2/2} \) almost everywhere (a.e.) for any \( |\lambda| < 1/\alpha_k \) and \( \nu_k, \alpha_k > 0 \). Then \( \sum_{k=1}^n D_k \) is sub-exponential with parameters \( (\sqrt{\sum_{k=1}^n \nu_k^2}, \alpha^*) \).

**Proof:** For \( \lambda \in (-1/\alpha^*, 1/\alpha^*) \), apply iterated expectation to get
\[
\mathbb{E} \left[ e^{\lambda (\sum_{k=1}^n D_k)} \right] = \mathbb{E} \left[ e^{\lambda (\sum_{k=1}^{n-1} D_k)} \mathbb{E} \left[ e^{\lambda D_n} \mid \mathcal{F}_{n-1} \right] \right] \\
\leq \mathbb{E} \left[ e^{\lambda \sum_{k=1}^{n-1} D_k} \right] e^{\lambda^2 \nu_n^2/2} \\
\leq e^{\lambda^2 \sum_{k=1}^n \nu_k^2/2}
\]
which proves the result. 

The sub-exponential tail bound provides the following inequality.

**Corollary 5.2**

\[
P \left[ \left| \sum_{k=1}^n D_k \right| \geq t \right] \leq \begin{cases} 
2e^{-\frac{t^2}{2\sum_{k=1}^n \nu_k^2}} & \text{if } 0 \leq t \leq \frac{\sum_{k=1}^n \nu_k^2}{\alpha^*} \\
2e^{-\frac{t}{\alpha^*}} & \text{if } t > \frac{\sum_{k=1}^n \nu_k^2}{\alpha^*}
\end{cases}
\]

Remember that bounded random variables are sub-Gaussian, which gives the following corollary.

**Corollary 5.3 [Azuma-Hoeffding]** Let \( \{D_k, \mathcal{F}_k\}_{k=1}^\infty \) be a martingale difference sequence such that \( D_k \in [a_k, b_k] \) almost surely for all \( k = 1, \ldots, n \). Then for all \( t > 0 \),
\[
P \left[ \left| \sum_{k=1}^n D_k \right| \geq t \right] \leq 2 \exp \left( -\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2} \right).
\]
Proof: Since $D_k \in [a_k, b_k]$ almost surely, the conditioned variable $(D_k|F_{k-1})$ is also bounded in $[a_k, b_k]$ almost surely. Therefore, $(D_k|F_{k-1})$ is sub-Gaussian at most $\sigma = (b_k - a_k)/2$ for all $k = 1, \ldots, n$. The result follows by Theorem 5.1 and Corollary 5.2 with parameters $(\sqrt{\sum_{k=1}^{n} (b_k - a_k)^2}/4, 0)$.

As an application of these results, we will establish a useful inequality, which is called the bounded difference inequality or McDiarmid’s inequality. Let us begin by defining the bounded difference property.

**Definition 5.4 [Bounded difference property]** A function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the bounded difference property (BDP) if there exists positive constants $(L_1, \ldots, L_n)$ such that for each $k = 1, 2, \ldots, n$,

$$|f(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, x', x_{k+1}, \ldots, x_n)| \leq L_k \quad \text{for all} \quad x, x' \in \mathbb{R}^d.$$

**Theorem 5.5 [Bounded difference inequality]** Suppose that $Z = f(X)$ satisfies the bounded difference property with parameters $(L_1, \ldots, L_n)$ and that the random vector $X = (X_1, \ldots, X_n)$ has independent elements. Then

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq t) \leq 2 \exp \left( - \frac{2t^2}{\sum_{k=1}^{n} L_k^2} \right) \quad \text{for all} \quad t \geq 0.$$

Proof: Start by constructing a martingale difference using the Doob martingale decomposition of $Z$ as

$$D_0 = \mathbb{E}(Z)$$
$$D_k = \mathbb{E}(Z|X_1, \ldots, X_k) - \mathbb{E}(Z|X_1, \ldots, X_{k-1}) \quad \text{for} \quad k = 1, \ldots, n.$$ 

Then we have $Z - \mathbb{E}(Z) = \sum_{k=1}^{n} D_k$. Define the random variables

$$A_k = \inf_x \mathbb{E}(Z|X_1, \ldots, X_{k-1}, x) - \mathbb{E}(Z|X_1, \ldots, X_{k-1}) \quad \text{and}$$
$$B_k = \sup_x \mathbb{E}(Z|X_1, \ldots, X_{k-1}, x) - \mathbb{E}(Z|X_1, \ldots, X_{k-1})$$

so that $B_k \geq A_k$ a.e. for all $k = 1, \ldots, n$. In addition,

$$D_k - A_k = \mathbb{E}(Z|X_1, \ldots, X_k) - \inf_x \mathbb{E}(Z|X_1, \ldots, X_{k-1}, x) \geq 0 \quad \text{a.e.}$$

Similarly, $B_k - D_k \geq 0$ a.e. Now observe that

$$D_k \leq B_k - A_k$$
$$\leq \sup_{x, x'} |\mathbb{E}[Z|X_1, \ldots, X_{k-1}, x] - \mathbb{E}[Z|X_1, \ldots, X_{k-1}, y]|$$
$$\leq L_k.$$

Apply the Azuma-Hoeffding inequality to get the result.

5.1.2 Applications

**Example 5.6 [Kernel density estimate]** Let $X_1, \ldots, X_n$ be independent and identically distributed random samples from a distribution $P$ with a Lebesgue-density $p = dP/d\mu$. We are interested in estimating the shape of $p$. Its kernel density estimate is

$$\hat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right) \quad \text{for} \quad x \in \mathbb{R},$$
where \( K(x) \geq 0, \int K(x)dx = 1 \) and \( h > 0 \). One way of measuring a proximity between \( \hat{p}_h \) and \( p \) is
\[
Z = \int_{-\infty}^{\infty} |\hat{p}_h(x) - p(x)|dx = f(X_1, \ldots, X_n).
\]
Then, denote \( \hat{p}_h'(x) \) for the kernel density estimate obtained by replacing \( X_i \) by \( X_i' \) and bound
\[
\left| f(X_1, \ldots, X_1', \ldots, X_n) - f(X_1, \ldots, X_i, \ldots, X_n) \right| = \left| \int_{-\infty}^{\infty} \hat{p}_h'(x) - p(x)dx - \int_{-\infty}^{\infty} \hat{p}_h(x) - p(x)dx \right|
\leq \frac{1}{nh} \int_{-\infty}^{\infty} |K\left(\frac{x - X_i'}{h}\right) - K\left(\frac{x - X_i}{h}\right)|dx
\leq \frac{1}{nh} \left[ h \int_{-\infty}^{\infty} K(z')dz' + h \int_{-\infty}^{\infty} K(z)dz \right] = \frac{2}{n}
\]
where we used the triangle inequality and the variable transformation to get the bound. This shows that \( f \) satisfies the bounded difference property with \( L_k = 2/n \) for all \( k = 1, \ldots, n \). Then McDiarmid’s inequality gives
\[
\mathbb{P}(|Z - \mathbb{E}(Z)| \geq t) \leq 2 \exp\left(-\frac{nt^2}{2}\right)
\]
where the upper bound does not depend on \( h \).

**Example 5.7 [Empirical measure]** Let \( A \) be a class of sets in \( \mathbb{R}^d \) and \( X_1, \ldots, X_n \) be independent and identically distributed random samples from a distribution \( \mathbb{P} \) on \( \mathbb{R}^d \). We are interested in
\[
Z = \sup_{A \in A} |\mathbb{P}(A) - \mathbb{P}_n(A)|
\]
where \( \mathbb{P}_n(A) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in A) \) is the empirical measure of \( A \). The empirical distribution function provides an example of empirical measures when \( d = 1 \). For a class \( A = \{(-\infty, x] : x \in \mathbb{R}\} \),
\[
Z_1 = \sup_t |F_n(t) - F(t)|
\]
where \( F_n(t) = \mathbb{P}_n((-\infty, t]) \) and \( F(t) = \mathbb{P}(X \leq t) \). In particular, Glivenko-Cantelli theorem says that \( Z_1 \to 0 \) almost surely. Later on, we will look into bounds on \( Z \). For now, denote \( Z = f(X_1, \ldots, X_n) \) and \( \mathbb{P}'_n(A) \) for the empirical measure of \( A \) obtained by replacing \( X_i \) by \( X_i' \)
\[
\left| f(X_1, \ldots, X_1', \ldots, X_n) - f(X_1, \ldots, X_i, \ldots, X_n) \right| = \sup_{A \in A} |\mathbb{P}(A) - \mathbb{P}_n(A)| - \sup_{A \in A} |\mathbb{P}(A) - \mathbb{P}_n(A)|
\leq \sup_{A \in A} |\mathbb{P}'_n(A) - \mathbb{P}_n(A)| = \frac{1}{n},
\]
Hence, \( Z \) satisfies the bounded difference property with \( L_k = 1/n \) for all \( k = 1, \ldots, n \). Then McDiarmid’s inequality provides
\[
\mathbb{P}(|Z - \mathbb{E}(Z)| \geq t) \leq 2 \exp(-2nt^2).
\]

### 5.2 Lipschitz functions of Gaussian variables

We investigate the concentration properties of Lipschitz functions of Gaussian variables. Let us say that a function \( f : \mathbb{R}^d \to \mathbb{R} \) is \( L \)-Lipschitz with respect to the Euclidean norm \( \| \cdot \|_2 \) if
\[
|f(x) - f(y)| \leq L\|x - y\|_2 \quad \text{for } x, y \in \mathbb{R}^d.
\]
A Lipschitz function is absolutely continuous and thus is differentiable almost everywhere. Now, the following theorem guarantees that any Lipschitz function of Gaussian variables is sub-Gaussian with parameter at most \( L \).

**Theorem 5.8** Let \((X_1, \ldots, X_n)\) be a vector of i.i.d. Gaussian variables from \( N(0, \sigma^2) \) and let \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) be \( L \)-Lipschitz. Then the variable \( f(X) - \mathbb{E}(f(X)) \) is sub-Gaussian with parameter at most \( L \), and thus

\[
P \left[ |f(X) - \mathbb{E}(f(X))| \geq t \right] \leq 2 \exp \left( -\frac{t^2}{2L^2\sigma^2} \right) \text{ for all } t \geq 0.
\]

Remarkably, this is a dimension free inequality.

**Proof:** Refer to [BLM13] in p.125. \( \blacksquare \)

**Example 5.9** [Maximum of Gaussian variables] For a random vector \( X = (X_1, \ldots, X_d) \sim N_d(0, \Sigma) \), define \( Z = \max_{1 \leq i \leq d} |X_i| \) and \( \sigma_{\text{max}}^2 = \max_{1 \leq i,j \leq d} \Sigma_{i,j} \). Then,

\[
P \left[ |Z - \mathbb{E}(Z)| \geq t \right] \leq 2 \exp \left( -\frac{t^2}{2\sigma_{\text{max}}^2} \right).
\]

**Proof:** Denote \( X = AW \) where \( W \sim N_d(0, I) \) and \( AA^T = \Sigma \). Then \( Z = \max_{1 \leq i \leq d} X_i = f(W) \) where \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is the function

\[
f(x) = \max_{1 \leq i \leq d} (Ax)_i
\]

Notice that the function \( f \) is Lipschitz with the parameter \( L = \max_{1 \leq i \leq d} \sqrt{\sum_{j=1}^d A_{i,j}^2} \) because for \( x, y \in \mathbb{R}^d \) we have

\[
|(Ax)_i - (Ay)_i| = \left| \sum_{j=1}^d A_{i,j} (x_j - y_j) \right| \\
\leq \sqrt{\sum_{j=1}^d A_{i,j}^2} \|x - y\|_2
\]

by Cauchy–Schwarz inequality. Furthermore,

\[
\sum_{j=1}^d A_{i,j}^2 = \mathbb{V}(X_i) = \mathbb{V} \left[ \sum_{j=1}^d A_{i,j} Z_j \right].
\]

Therefore, \( f \) is \( \sigma_{\text{max}} \)-Lipschitz. The proof is done by Theorem 5.8. \( \blacksquare \)

### 5.3 Covering and packing number

Let \( Y_i \) be \( X_i \) or \( |X_i| \) where \( X_i \) is sub-Gaussian or sub-Exponential. In this case, we are often interested in \( \max_{i \in I} Y_i \) or \( \mathbb{E} \left[ \max_{i \in I} Y_i \right] \) for a given class \( I \). If the size of \( I \) is infinite, it is challenging to develop uniform bounds. To tackle this problem, we will discretize \( I \) by picking a finite subset \( \tilde{I} \) of \( I \) and then approximating \( \max_{i \in I} Y_i \) with \( \max_{i \in \tilde{I}} Y_i \). Before we go into the details, let us define a metric space.
Definition 5.10 [Metric space] A metric space is an ordered pair \((X, d)\) where \(X\) is a set and \(d\) is a metric on \(X\) such that \(d: X \times X \to \mathbb{R}\) and for \(x, y, z \in X\), the following holds.

1. \(d(x, y) \geq 0\) and \(d(x, y) = 0\) if and only if \(x = y\)
2. \(d(x, y) = d(y, x)\)
3. \(d(x, z) \leq d(x, y) + d(y, z)\)

Example 5.11 Here are some examples of metric spaces.

- \((\mathbb{R}^d, ||\cdot||)\) and \(||x|| = \sqrt{\sum_i x_i^2}\)
- \((\mathbb{R}^d, ||\cdot||_p)\) and \(||x||_p = (\sum_i x_i^p)^{1/p}\) for \(p \geq 1\)
- \((\mathbb{R}^d, ||\cdot||_\infty)\) and \(||x||_\infty = \max_i |x_i|\)
- \((\{0,1\}^d, d_H)\) and \(d_H(x, y) = \frac{1}{d} \sum_{i=1}^{d} I(x_i \neq y_i)\) called the Hamming distance

Lastly, we talked about \(L_p\)-space. Let \(X = \{f : [0,1] \to \mathbb{R}\}\) be a set of functions. An \(L_p\)-space on \([0,1]\) contains functions of \(X\) for which the \(p\)-th power of the absolute value is \(\mu\)-integrable. That is

\[
||f||_p = \left( \int_0^1 |f|^p d\mu \right)^{1/p} < \infty
\]

where \(\mu\) is a measure on \([0,1]\) and \(p \geq 1\). The common choice of \(p\) is \(p = 2\), which allows a richer theory. The \(L_p\)-distance between \(f\) and \(g\) is defined as

\[
||f - g||_p = \left( \int_0^1 |f(x) - g(x)|^p d\mu \right)^{\frac{1}{p}}.
\]

Especially, when \(p = \infty\),

\[
||f - g||_\infty = \sup_{x \in [0,1]} |f(x) - g(x)|.
\]

References