Martingale Sequence Review

Definition. A sequence \( \{Y_n\}_{n=1}^{\infty} \) is a martingale sequence w.r.t. \( \{X_n\}_{n=1}^{\infty} \) if

\[
\begin{align*}
- & Y_n \text{ is a measurable function of } X_1, \cdots, X_n; \\
- & E[|Y_n|] < \infty, \ \forall n; \\
- & E[Y_{k+1}|X_1, \cdots, X_k] = Y_k, \ \forall k.
\end{align*}
\]

Examples.

1. \( Y_k = E[f(X)|X_1, \cdots, X_k] \) is a martingale given \( E[|f(X)|] < \infty \).

2. \( \{X_n\}_{n=1}^{\infty} \) is a sequence of 0-mean independent RV’s. If \( S_n = \sum_{i=1}^{n} X_i \), then \( \{S_n\}_{n=1}^{\infty} \) is a martingale.

Proof: \( S_n \) satisfies the 3 conditions of the definition of martingales.

\[
\begin{align*}
- & S_n \text{ is a partial sum of } \{X_i\}_{i=1}^{n}, \text{ so it’s measurable.} \\
- & E[|S_n|] \leq \sum_{i=1}^{n} E[|X_i|] < \infty. \\
- & E[S_{n+1}|X_1, \cdots, X_n] = S_n, \text{ because} \\
& \quad E[S_{n+1}|X_1, \cdots, X_n] = E[S_n + X_{n+1}|X_1, \cdots, X_n] \\
& \quad = S_n + E[X_{n+1}|X_1, \cdots, X_n] \quad S_n \text{ is a constant conditioning on } X_1, \cdots, X_n \\
& \quad = S_n + E[X_{n+1}] \quad X_{n+1} \text{ is independent of } X_1, \cdots, X_n \\
& \quad = S_n \quad X_{n+1} \text{ has zero-mean}
\end{align*}
\]

7.1 Martingale Difference Sequence

Definition 7.1. \( \{D_k\}_{k=1}^{\infty} \) is a martingale difference sequence (abbr. MDS) w.r.t. \( \{X_k\}_{k=1}^{\infty} \) if

\[
\begin{align*}
- & D_k \text{ is a measurable function of } X_1, \cdots, X_k;
\end{align*}
\]
- \(\mathbb{E}[|D_k|] < \infty, \forall k\);
- \(\mathbb{E}[D_{k+1}|X_1,\cdots,X_k] = 0, \forall k\).

**Example.** Suppose \(\{Y_k\}_{k=1}^{\infty}\) is a martingale sequence w.r.t. \(\{X_k\}_{k=1}^{\infty}\). Let \(D_k = Y_k - Y_{k-1}, k = 2,3,\ldots\)

- \(D_k\) is measurable because \(Y_k, Y_{k-1}\) are measurable.
- \(\mathbb{E}[|D_k|] \leq \mathbb{E}[|Y_k|] + \mathbb{E}[|Y_{k-1}|] < \infty\).
- \(\mathbb{E}[D_{k+1}|X_1,\cdots,X_n] = D_k\), because

\[
\mathbb{E}[D_{k+1}|X_1,\cdots,X_k] = \mathbb{E}[Y_{k+1} - Y_k|X_1,\cdots,X_k]
= \mathbb{E}[Y_{k+1}|X_1,\cdots,X_k] - Y_k
= 0
\]

\(Y_k\) is a constant conditioning on \(X_1,\cdots,X_k\)

\(Y\) is a martingale, so it equals \(Y_k - Y_{k-1}\)

Hence \(\{D_k\}_{k=1}^{\infty}\) is a MDS w.r.t. \(\{X_k\}_{k=1}^{\infty}\). Note that \(Y_n - Y_0 = \sum_{k=1}^{n} D_k\).

**Theorem 7.2.** Suppose \(\{D_k\}_{k=1}^{\infty}\) is a MDS w.r.t. \(\{X_k\}_{k=1}^{\infty}\), satisfying

\[
\mathbb{E}[e^{\lambda D_n}|X_1,\cdots,X_{n-1}] \leq \exp\left(\frac{\lambda^2 \nu_n^2}{2}\right), \quad \forall \lambda \in \left[0, \frac{1}{\alpha_n}\right],
\]

i.e. \(D_n|X_1,\cdots,X_{n-1} \sim \text{SE}(\nu_n, \alpha_n)\). Define \(\nu_n^* = \sqrt{\nu_1^2 + \cdots + \nu_n^2}, \alpha_n^* = \max_{k=1}^{n} \alpha_k\). Then,

\[
\sum_{k=1}^{n} D_k \sim \text{SE}\left(\nu_n^*, \alpha_n^*\right) \implies \mathbb{P}\left\{ \sum_{k=1}^{n} D_k > t \right\} \leq \exp\left(-\frac{t^2}{2\nu_n^*}\right), \quad \forall t \in \left[0, \frac{1}{\alpha_n^*}\right].
\]

**Proof:**

\[
\mathbb{E}_{X_1,\ldots,n}\left[\exp\left(\lambda \sum_{k=1}^{n} D_k\right)\right] = \mathbb{E}_{X_1,\ldots,n-1}\left[\mathbb{E}_{X_n}\left[\exp\left(\lambda \sum_{k=1}^{n} D_k\right)\right]\right]
= \mathbb{E}_{X_1,\ldots,n-1}\left[\exp\left(\lambda \sum_{k=1}^{n-1} D_k\right)\mathbb{E}_{X_n}\left[\exp\left(\lambda D_n\right)|X_1,\cdots,X_{n-1}\right]\right], \quad \forall \lambda \in \left[0, \frac{1}{\alpha_n}\right]
\leq \exp\left(\frac{\lambda^2 \nu_n^2}{2}\right) \mathbb{E}_{X_1,\ldots,n-1}\left[\exp\left(\lambda \sum_{k=1}^{n-1} D_k\right)\right]
\leq \cdots \leq \exp\left(\frac{\lambda^2}{2} \sum_{k=1}^{n} \nu_k^2\right), \quad \forall \lambda \in \bigcap_{k=1}^{n} \left[0, \frac{1}{\alpha_k}\right] = \left[0, \frac{1}{\max_{k=1}^{n} \alpha_k}\right].
\]

Azuma Hoeffding]

**Theorem 7.3 (Azuma-Hoeffding Inequality).** For a sequence of Martingale Difference Sequence random variable \(\{D_k\}_{k=1}^{\infty}\) with respect to some other sequence of random variable \(\{X_k\}_{k=1}^{\infty}\), if we have \(D_k \in [a_k, b_k]\) almost sure for some constant \(a_k, b_k\) and \(k = 1,2,\ldots,n\), Then:

\[
\mathbb{P}\left(\sum_{k=1}^{n} D_k > t\right) \leq e^{-\frac{t^2}{\sum_{k=1}^{n} (b_k - a_k)^2}}
\]
Proof: Recall that by Hoeffding’s lemma \([?]\), we have that \(D_k \sim SG\left(\frac{b_k - a_k}{2}\right)\), we have that

\[
D_k \mid X_1, \ldots, X_{k-1} \sim SG\left(b_k - a_k^2\right),
\]

\[
E\left[e^{\lambda \sum_{k=1}^{n} D_k}\right] = E_{X_1, \ldots, X_{n-1}} \left[ E_{X_n} \left[ \exp \left( \lambda \sum_{k=1}^{n} D_k \right) \mid X_1, \ldots, X_{n-1} \right] \right]
\]

\[
= E_{X_1, \ldots, X_{n-1}} \left[ E_{X_n} \left[ \exp \left( \lambda \sum_{k=1}^{n-1} D_k \right) \exp \left( \lambda D_n \right) \mid X_1, \ldots, X_{n-1} \right] \right]
\]

\[
\leq E_{X_1, \ldots, X_{n-1}} \left[ \exp \left( \lambda \sum_{k=1}^{n-1} D_k \right) \exp \left( \frac{\lambda^2 (b_k - a_k)^2}{8} \right) \right]
\]

\[
= \exp \left( \frac{\lambda^2 (b_k - a_k)^2}{8} \right) E_{X_1, \ldots, X_{n-1}} \left[ \exp \left( \lambda \sum_{k=1}^{n-1} D_k \right) \right]
\]

By iteratively derive the bound we could get that:

\[
E\left[e^{\lambda \sum_{k=1}^{n} D_k}\right] \leq e^{\frac{\lambda^2 \sum_{k=1}^{n} (b_k - a_k)^2}{8}}
\]

That is \(\sum_{k=1}^{n} D_k \sim SG\left(\frac{1}{2} \sqrt{\sum_{k=1}^{n} (b_k - a_k)^2}\right)\), By that we can prove that:

\[
P\left(\sum_{k=1}^{n} D_k > t\right) \leq e^{-\frac{2 t^2}{\sum_{k=1}^{n} (b_k - a_k)^2}}
\]

Recall that a sequence of random variable \(\{Y_k\}_{k=1}^{\infty}\) where \(Y_k = E[f(x)\mid X_1, \ldots, X_n]\) respect to some sequence of random variable \(\{X_k\}_{k=1}^{\infty}\) is a Martingale sequence, then the sequence of \(\{D_k\}_{k=1}^{\infty}\) where \(D_k = Y_k - Y_{k-1}\) is a Martingale Difference Sequence. We have that:

\[
Y_n - Y_0 = \sum_{k=1}^{n} D_k
\]

Where \(Y_n = f(x)\) and \(Y_0 = E[f(x)]\), under this condition, we can bound the ERM with Azuma-Hoeffding Inequality.

### 7.2 Bounded Difference Inequality

**Theorem 7.4 (Bounded Difference Inequality).** Let \(X_1, \ldots, X_n\) be a set of random variables, \(f : \mathbb{R}^n \rightarrow \mathbb{R}\), if for all \(k \in \{1, 2, \ldots, n\}\), we have a set of constant \(L_k\) where:

\[
|f(X_1, \ldots, X_k, \ldots, X_n) - f(X_1, \ldots, X'_k, \ldots, X_n)| \leq L_k
\]

Then we have the following equation:

\[
P(|f(x) - E[f(x)]| > t) \leq 2e^{-\frac{2 t^2}{\sum_{k=1}^{n} L_k^2}}
\]
**Proof:** Consider a sequence of random variable \( \{D_k\}_{k=1}^{\infty} \) where \( D_k = \mathbb{E}[f(x)|X_1, \ldots, X_k] - \mathbb{E}[f(x)|X_1, \ldots, X_{k-1}] \). We first proof that \( D_k \sim SG(\frac{L_k}{2}) \). Denote \( B_k \) and \( A_k \) as the following:

\[
A_k = \inf_x \mathbb{E}[f(x)|X_1, \ldots, X_{k-1}, X] - \mathbb{E}[f(x)|X_1, \ldots, X_{k-1}]
\]
\[
B_k = \sup_x \mathbb{E}[f(x)|X_1, \ldots, X_{k-1}, X] - \mathbb{E}[f(x)|X_1, \ldots, X_{k-1}]
\]

we have:

\[
D_k - A_k = \mathbb{E}[f(x)|X_1, \ldots, X_k] - \inf_x \mathbb{E}[f(x)|X_1, \ldots, X_{k-1}, X] \geq 0
\]
\[
B_k - D_k = \sup_x \mathbb{E}[f(x)|X_1, \ldots, X_k] - \mathbb{E}[f(x)|X_1, \ldots, X_{k-1}] \geq 0
\]

That is \( A_k \leq D_k \leq B_k \) almost surely.

\[
B_k - A_k = \sup_x \mathbb{E}[f(x)|X_1, \ldots, X_{k-1}, X] - \inf_y \mathbb{E}[f(x)|X_1, \ldots, X_{k-1}, Y]
\]
\[
= \sup_x (\mathbb{E}[f(x)|X_1, \ldots, X_{k-1}, X] - \mathbb{E}[f(x)|X_1, \ldots, X_{k-1}, Y])
\]
\[
\leq L_k
\]

That is \( D_k \sim SG(\frac{L_k}{2}) \).

By the Asuma-Hoeffding Inequality prove we get \( \sum_{k=1}^{n} D_k \sim SG(\frac{1}{2} \sqrt{\sum_{k=1}^{n} L_k^2}) \), which result in:

\[
\mathbb{P}(|f(x) - \mathbb{E}[f(x)]| > t) \leq 2e^{-\frac{t^2}{\sum_{k=1}^{n} L_k^2}}
\]

Bounded Difference Inequality theorem is very powerful in that it can calculate the tailbounds for functions of non-independent random variables.

**Example:** Let \( f(x_1, \ldots, x_n) = \sum_{i=1}^{n} (x_i - \mu_i) \) where \( x_i \in [a_i, b_i] \), we have:

\[
|f(x_1, \ldots, x_k, \ldots, x_n) - f(x_1, \ldots, x_k', \ldots, x_n)| = |x_k - x_k'| \leq b_k - a_k
\]

By using Bounded Difference Inequality we get:

\[
\mathbb{P}(|f(x) - \mathbb{E}[f(x)]| > t) \leq e^{-\frac{t^2}{\sum_{k=1}^{n} (b_k - a_k)^2}}
\]

**Example:** U statistics

Define a function \( f \) on \( \{X_k\}_{k=1}^{\infty} : f(X_1, \ldots, X_n) = \frac{1}{\binom{n}{2}} \sum_{i<j} g(X_i, X_j) \) where \( g : \mathbb{R}^2 \rightarrow \mathbb{R} \) is a symbolic function and \( g(x, y) \leq b, \forall x, y \). We can prove that \( f \) satisfies Bounded Difference Inequality.

**Proof:**

\[
f(X_1, \ldots, X_k, \ldots, X_n) - f(X_1, \ldots, X_k', \ldots, X_n) = \frac{1}{\binom{n}{2}} \sum_{j \neq k} g(X_j, X_k) - g(X_j, X_k')
\]
\[
\leq \frac{2(2b)}{n(n-1)} \leq \frac{4b}{n}
\]
As a result, plugging it into Bounded Difference Inequality where \( L_k = \frac{4b}{n} \), we get:

\[
P(|f(X) - \mathbb{E}[f(X)]| > t) \leq \exp \left( -\frac{2t^2}{n(\frac{4b}{n})^2} \right) = \exp \left( -\frac{2nt^2}{8b^2} \right)
\]

**Example: Rademacher Complexity**

If \( \epsilon_1...\epsilon_n \) are Rademacher random variables where \( \epsilon_n \in [-1, +1] \) with equal probabilities. Then we define a function \( f(\epsilon_1...\epsilon_n) = R_n(A) = \sup_{a \in A} a^T \epsilon (A \subseteq \mathbb{R}^n) \) and it satisfies Bounded Difference Inequality.

**Proof:**

\[
f(\epsilon_1...\epsilon_k...\epsilon_n) - f(\epsilon_1...\epsilon_k'...\epsilon_n) \leq \sup_{a \in A} a^T \epsilon - \sup_{a \in A} a^T \epsilon'
\]

\[
\leq \langle a^*, \epsilon \rangle - \langle a^*, \epsilon' \rangle \quad (a^* = \sup_{a \in A} a^T \epsilon)
\]

\[
\leq \langle a^*, \epsilon - \epsilon' \rangle
\]

\[
= a_k^* (\epsilon_k - \epsilon_k')
\]

\[
\leq 2|a_k^*| \leq 2 \sup_{a_k} |a_k|
\]

As a result, plugging it into Bounded Difference Inequality where \( L_k = \sup_{a_k} |a_k| \), we get:

\[
f(\epsilon) - \mathbb{E}[f(\epsilon)] = R_n(A) - \mathbb{E}[R_n(A)] \sim \mathcal{SG} \left( \sqrt{\sum_{k=1}^{n} \sup_{a \in A} |a_k|^2} \right)
\]

**Example: Lipschitz functions**

We can bound \( |f(x) - f(y)|(x,y) \) only differs in \( k^{th} \) coordinate) by the distance between \( x \) and \( y \) according to some distance metric if \( f \) satisfies Lipschitz conditions. For example, if \( f \) is Lipschitz w.r.t. Hamming distance, then

\[
|f(x) - f(y)| \leq L \cdot d_H(x, y) = L \cdot \sum_{i=1}^{n} I(x_i \neq y_i)
\]

**Theorem 7.5.** If \( X_1,...,X_n, \text{ iid, is stand Gaussian with distribution } N(0,1) \) and \( f \) is \( L_n \)-Lipschitz w.r.t. \( L_2 \)-norm distance, i.e, \( |f(x) - f(y)| \leq L_n \cdot \|x - y\|_2, \forall x, y \in \mathbb{R}^n \) Then:

\[
P(|f(x) - \mathbb{E}[f(x)]| > t) \leq 2 \exp \left( -\frac{t^2}{2L_n^2} \right)
\]

The proof is very hard and will be omitted. For example, if \( X_1...X_n, \text{ iid, is stand Gaussian with distribution } N(0,1) \) and \( X(1),...,X(n) \) is a function of \( X_1,...,X_n \) that it orders it such that \( X(1) \geq X(2),...,\geq X(k),...,\geq X(n) \) where \( X(k) \) is the \( k^{th} \) largest. Then, if we \( X(n) \) and \( Y(n) \) only differs in \( k^{th} \) component, according to the pigeonhole principle, we have:

\[
|X(k) - Y(k)| \leq \|X - Y\|_2
\]

As a result:

\[
P(|X(k) - \mathbb{E}[X(k)]| > t) \leq 2 \exp \left( -\frac{t^2}{2} \right)
\]
**Example: Gaussian Complexity**

$X_1, \ldots, X_n$, iid, is standard Gaussian with distribution $N(0, 1)$. $R(A) = \sup_{a \in A} \langle a, X \rangle$ with $A \in \mathbb{R}^n$ and $f(X) = R_n(A)$ and $X, Y$ only differs in the $k$th coordinate.

Then similar to the Rademacher Complexity example, we have:

$$f(X) - f(Y) \leq \langle a^*, X - Y \rangle$$

$$(a^* = \sup_{a \in A} \| a, X \|)$$

$$\leq \| a^* \|_2 \| X - Y \|_2$$

Cauchy Schwartz Inequality

$$\leq \sup_{a \in A} \| a \|_2 \| X - Y \|_2$$

As a result, applying Bounded Difference Inequality:

$$f(X) - \mathbb{E}[f(X)] \sim SG(\sup_{a \in A} \| a \|_2)$$