10-716: Advanced Machine Learning	**Spring 2019**

Lecture 7: February 7

Lecturer: Pradeep Ravikumar *Scribes: Jiamei Liu, Kaiji Lu, Weizhao Tang*

Martingale Sequence Review

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

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

Examples.

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

   - S_n is a partial sum of $\{X_i\}_{i=1}^n$, so it's measurable.
   - $\mathbb{E}[|S_n|] \le \sum_{i=1}^n \mathbb{E}[|X_i|] < \infty$.
   - $\mathbb{E}\left[S_{n+1}|X_1, \cdots X_n\right] = S_n$, because

$$\begin{aligned}
\mathbb{E}\left[S_{n+1}|X_1, \cdots X_n\right] &= \mathbb{E}\left[S_n + X_{n+1}|X_1, \cdots X_n\right] \\
&= S_n + \mathbb{E}\left[X_{n+1}|X_1, \cdots X_n\right] && S_n \text{ is a constant conditioning on } X_1, \cdots, X_n \\
&= S_n + \mathbb{E}[X_{n+1}] && X_{n+1} \text{ is independent of } X_1, \cdots X_n \\
&= S_n && X_{n+1} \text{ has zero-mean}
\end{aligned}$$

■

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*

- D_k *is a measurable function of* $X_1, \cdots, X_k$*;*

- $\mathbb{E}[|D_k|] < \infty,\ \forall k$;
- $\mathbb{E}[D_{k+1}|X_1, \cdots, X_k] = 0,\ \forall k$.

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

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

$$\begin{aligned}
\mathbb{E}\left[D_{k+1}|X_1, \cdots X_k\right] &= \mathbb{E}\left[Y_{k+1} - Y_k|X_1, \cdots X_n\right] \\
&= \mathbb{E}\left[Y_{k+1}|X_1, \cdots X_k\right] - Y_k && Y_k \text{ is a constant conditioning on } X_1, \cdots, X_k \\
&= 0 && Y \text{ is a martingale, so it equals } Y_k - Y_k
\end{aligned}$$

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*

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

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

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

Proof:

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

■

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 $\{X_n\}_{k=1}^{\infty}$, 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:*

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

Proof: Recall that by hoeffding's lemma [?] $D_k \sim SG(\frac{b_k - a_k}{2})$, we have that $D_k|X_1, \ldots X_k - 1 \sim SG(\frac{b_k - a_k}{2})$,

$$\begin{aligned}
\mathbb{E}[e^{\lambda \sum_{k=1}^{n} D_k}] &= \mathbb{E}_{X_1, \ldots X_{n-1}}\left[\mathbb{E}_{X_n}[\exp(\lambda \sum_{k=1}^{n} D_k)|X_1, \ldots, X_{n-1}]\right] \\
&= \mathbb{E}_{X_1, \ldots, X_{n-1}}\left[\mathbb{E}_{X_n}[\exp(\lambda \sum_{k=1}^{n-1} D_k)\exp(\lambda D_n)|X_1, \ldots, X_{n-1}]\right] \\
&= \mathbb{E}_{X_1, \ldots, X_{n-1}}\left[\exp(\lambda \sum_{k=1}^{n-1} D_k)\mathbb{E}_{X_n}[\exp(\lambda D_n)|X_1, \ldots, X_{n-1}]\right] \\
&\leq \mathbb{E}_{X_1, \ldots, X_{n-1}}\left[\exp(\lambda \sum_{k=1}^{n-1} D_k)\exp(\frac{\lambda^2 (b_k - a_k)^2}{8})\right] \\
&= \exp(\frac{\lambda^2 (b_k - a_k)^2}{8})\mathbb{E}_{X_1, \ldots, X_{n-1}}[\exp(\lambda \sum_{k=1}^{n-1} D_k)]
\end{aligned}$$

By iteratively derive the bound we could get that:

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

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

$$\mathbb{P}(\sum_{k=1}^{n} D_k > t) \leq e^{\frac{-2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}}$$

■

Recall that a sequence of random variable $\{Y_k\}_{k=1}^{\infty}$ where $Y_k = \mathbb{E}[f(x)|X_1, \ldots, X_n]$ 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 $Y_0 = \mathbb{E}[f(x)]$, 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, $f : \mathbb{R}^n \to \mathbb{R}$, 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:

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

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

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

we have:

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

That is $A_k \leq D_k \leq B_k$ almost surely.

$$\begin{aligned} B_k - A_k &= \sup_x \mathbb{E}[f(x)|X_1,\dots,X_{k-1},X] - \inf_y \mathbb{E}[f(x)|X_1,\dots,X_{k-1},Y] \\ &= \sup_{x,y}(\mathbb{E}[f(x)|X_1,\dots,X_{k-1},X] - \mathbb{E}[f(x)|X_1,\dots,X_{k-1},Y]) \\ &\leq L_k \end{aligned}$$

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:

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

■

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

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

$$|f(x_1,\dots,x_k,\dots,x_n) - f(x_1,\dots,x_k',\dots,x_n)| = |x_k - x_k'| \leq b_k - a_k$$

By using Bounded Difference Inequality we get:

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

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

Proof:

$$\begin{aligned} f(X_1,...,X_k,...,X_n) - f(X_1,...,X_k',...,X_n) &= \frac{1}{\binom{n}{2}} \sum_{j\neq k} g(X_j,X_k) - g(X_j,X_k') \\ &\leq \frac{2(2b)}{n(n-1)} \leq \frac{4b}{n} \end{aligned}$$

As a result, plugging it into Bounded Difference Inequality where $L_k = \frac{4b}{n}$, we get:

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

■

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 $f(\epsilon_1 ... \epsilon_n) = R_n(A) = \sup_{a \in A} a^T \boldsymbol{\epsilon} (A \subseteq \mathbb{R}^n)$ and it satisfies Bounded Difference Inequality.

Proof:

$$\begin{aligned}
f(\epsilon_1 ... \epsilon_k ... \epsilon_n) - f(\epsilon_1 ... \epsilon_k' ... \epsilon_n) &\leq \sup_{a \in A} a^T \boldsymbol{\epsilon} - \sup_{a \in A} a^T \bar{\boldsymbol{\epsilon}} \\
&\leq \langle a^*, \boldsymbol{\epsilon} \rangle - \langle a^*, \bar{\boldsymbol{\epsilon}} \rangle \qquad (a^* = \sup_{a \in A} a^T \boldsymbol{\epsilon}) \\
&\leq \langle a^*, \boldsymbol{\epsilon} - \bar{\boldsymbol{\epsilon}} \rangle \\
&= a_k^* (\epsilon_k - \epsilon_k') \\
&\leq 2|a_k^*| \leq 2 \sup_{a_k} |a_k|
\end{aligned}$$

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

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

■

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

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

Theorem 7.5. *If $X_1, ..., X_n$, 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:*

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

The proof is very hard and will be omitted. For example, if $X_1 ... X_n$, 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 $\boldsymbol{X_{(n)}}$ and $\boldsymbol{Y_{(n)}}$ only differs in k^{th} component, according to the pigeonhole principle, we have:

$$|X_{(k)} - Y_{(k)}| \leq \|X - Y\|_2$$

As a result:

$$\mathbb{P}(|X_{(k)} - \mathbb{E}[X_{(k)}]| > t) \leq 2\exp\left(\frac{-t^2}{2}\right)$$

Example: Gaussian Complexity
$X_1, ..., X_n$, iid, is stand 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:

$$\begin{aligned} f(X) - f(Y) &\leq \langle a^*, X - Y\rangle && (a^* = \sup_{a\in A}\|a, X\|) \\ &\leq \|a^*\|_2 \|X - Y\|_2 && \text{Cauchy Schwartz Inequality} \\ &\leq \sup_{a\in A}\|a\|_2 \|X - Y\|_2 \end{aligned}$$

As a result, applying Bounded Difference Inequality:

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