

# SDS 384 11: Theoretical Statistics

## Lecture 5: Martingale inequalities

---

Purnamrita Sarkar  
Department of Statistics and Data Science  
The University of Texas at Austin

## A bit background

- So far we have looked at sums of random variables. What if want to study properties of functions of independent random variables?

## A bit background

- So far we have looked at sums of random variables. What if want to study properties of functions of independent random variables?
- Consider  $n$  independent random variables  $X = (X_1, \dots, X_n)$ .

## A bit background

- So far we have looked at sums of random variables. What if want to study properties of functions of independent random variables?
- Consider  $n$  independent random variables  $X = (X_1, \dots, X_n)$ .
- We want to bound  $f(X_1, \dots, X_n) - E[f(X_1, \dots, X_n)]$

## A bit background

- So far we have looked at sums of random variables. What if want to study properties of functions of independent random variables?
- Consider  $n$  independent random variables  $X = (X_1, \dots, X_n)$ .
- We want to bound  $f(X_1, \dots, X_n) - E[f(X_1, \dots, X_n)]$
- Define  $Y_k = E[f(X)|X_1, \dots, X_k]$  for  $k \in \{1, \dots, n-1\}$

## A bit background

- So far we have looked at sums of random variables. What if want to study properties of functions of independent random variables?
- Consider  $n$  independent random variables  $X = (X_1, \dots, X_n)$ .
- We want to bound  $f(X_1, \dots, X_n) - E[f(X_1, \dots, X_n)]$
- Define  $Y_k = E[f(X)|X_1, \dots, X_k]$  for  $k \in \{1, \dots, n-1\}$
- $Y_0 = E[f(X)]$  and  $Y_n = f(X)$

## A bit background

- So far we have looked at sums of random variables. What if want to study properties of functions of independent random variables?
- Consider  $n$  independent random variables  $X = (X_1, \dots, X_n)$ .
- We want to bound  $f(X_1, \dots, X_n) - E[f(X_1, \dots, X_n)]$
- Define  $Y_k = E[f(X)|X_1, \dots, X_k]$  for  $k \in \{1, \dots, n-1\}$
- $Y_0 = E[f(X)]$  and  $Y_n = f(X)$
- Now  $f(X) - E[f(X)] = \sum_{i=0}^{n-1} \underbrace{(Y_{i+1} - Y_i)}_{D_i}$

## A bit background

- So far we have looked at sums of random variables. What if want to study properties of functions of independent random variables?
- Consider  $n$  independent random variables  $X = (X_1, \dots, X_n)$ .
- We want to bound  $f(X_1, \dots, X_n) - E[f(X_1, \dots, X_n)]$
- Define  $Y_k = E[f(X)|X_1, \dots, X_k]$  for  $k \in \{1, \dots, n-1\}$
- $Y_0 = E[f(X)]$  and  $Y_n = f(X)$
- Now  $f(X) - E[f(X)] = \sum_{i=0}^{n-1} \underbrace{(Y_{i+1} - Y_i)}_{D_i}$
- This forms a Martingale difference sequence.



# Martingales

## Definition

A sequence of random variables  $\{Y_i\}$  adapted to a filtration  $\mathcal{F}_i$  is a martingale if, for all  $i$ ,

$$E|Y_i| < \infty \quad E[Y_{i+1}|\mathcal{F}_i] = Y_i$$

- A filtration  $\{\mathcal{F}_i\}$  is a sequence of nested  $\sigma$ -fields, i.e.  $\mathcal{F}_i \subseteq \mathcal{F}_{i+1}$ .
- $Y_i$  is adapted to  $\mathcal{F}_i$  means that each  $Y_i$  is measurable w.r.t  $\mathcal{F}_i$ .

# Martingales

## Definition

A sequence of random variables  $\{Y_i\}$  adapted to a filtration  $\mathcal{F}_i$  is a martingale if, for all  $i$ ,

$$E|Y_i| < \infty \quad E[Y_{i+1}|\mathcal{F}_i] = Y_i$$

- A filtration  $\{\mathcal{F}_i\}$  is a sequence of nested  $\sigma$ -fields, i.e.  $\mathcal{F}_i \subseteq \mathcal{F}_{i+1}$ .
- $Y_i$  is adapted to  $\mathcal{F}_i$  means that each  $Y_i$  is measurable w.r.t  $\mathcal{F}_i$ .
- If  $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$ , then we say that  $\{Y_i\}$  forms a martingale sequence w.r.t  $\{X_i\}$ .

# Martingales

## Definition

A sequence of random variables  $\{Y_i\}$  adapted to a filtration  $\mathcal{F}_i$  is a martingale if, for all  $i$ ,

$$E|Y_i| < \infty \quad E[Y_{i+1}|\mathcal{F}_i] = Y_i$$

- A filtration  $\{\mathcal{F}_i\}$  is a sequence of nested  $\sigma$ -fields, i.e.  $\mathcal{F}_i \subseteq \mathcal{F}_{i+1}$ .
- $Y_i$  is adapted to  $\mathcal{F}_i$  means that each  $Y_i$  is measurable w.r.t  $\mathcal{F}_i$ .
- If  $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$ , then we say that  $\{Y_i\}$  forms a martingale sequence w.r.t  $\{X_i\}$ .
- If  $\mathcal{F}_i = \sigma(Y_1, \dots, Y_i)$ , then we say that  $\{Y_i\}$  forms a martingale sequence.

## Example-partial sums of i.i.d sequences

### Example

Let  $\{X_i\}_{i=1}^{\infty}$  be a sequence of i.i.d random variables with  $E[X_1] = \mu$ . Let  $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$ . Then  $\{Y_i = \sum_{k=1}^i X_k - i\mu\}$  is a martingale sequence w.r.t  $\{X_i\}$ .

## Example-partial sums of i.i.d sequences

### Example

Let  $\{X_i\}_{i=1}^{\infty}$  be a sequence of i.i.d random variables with  $E[X_1] = \mu$ . Let  $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$ . Then  $\{Y_i = \sum_{k=1}^i X_k - i\mu\}$  is a martingale sequence w.r.t  $\{X_i\}$ .

- $Y_i$  is measurable w.r.t  $\mathcal{F}_i$ .

## Example-partial sums of i.i.d sequences

### Example

Let  $\{X_i\}_{i=1}^{\infty}$  be a sequence of i.i.d random variables with  $E[X_1] = \mu$ . Let  $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$ . Then  $\{Y_i = \sum_{k=1}^i X_k - k\mu\}$  is a martingale sequence w.r.t  $\{X_i\}$ .

- $Y_i$  is measurable w.r.t  $\mathcal{F}_i$ .
- Finally,

$$\begin{aligned} E[Y_{i+1} | \mathcal{F}_i] &= E[X_{i+1} + \sum_{k=1}^i X_k - (i+1)\mu | \mathcal{F}_i] \\ &= \mu + \sum_{k=1}^i X_k - (i+1)\mu = Y_i \end{aligned}$$

## Example

Let  $\{X_i\}_{i=1}^{\infty}$  be a sequence of i.i.d random variables. Let  $Y_i = E[f(X)|X_1, \dots, X_i]$  and assume that  $E[|f(X)|] < \infty$ . Then  $\{Y_i\}_{i=0}^n$  is a martingale sequence w.r.t  $\{X_i\}_{i=1}^n$ .

## Example

Let  $\{X_i\}_{i=1}^{\infty}$  be a sequence of i.i.d random variables. Let  $Y_i = E[f(X)|X_1, \dots, X_i]$  and assume that  $E[|f(X)|] < \infty$ . Then  $\{Y_i\}_{i=0}^n$  is a martingale sequence w.r.t  $\{X_i\}_{i=1}^n$ .

- $E[|Y_i|] = E[|E[f(X)|X_1, \dots, X_i]|] \leq E[|f(X)|] < \infty$ . (Use Jensen on  $|(\cdot)|$ )



# Doob construction

## Example

Let  $\{X_i\}_{i=1}^{\infty}$  be a sequence of i.i.d random variables. Let  $Y_i = E[f(X)|X_1, \dots, X_i]$  and assume that  $E[|f(X)|] < \infty$ . Then  $\{Y_i\}_{i=0}^n$  is a martingale sequence w.r.t  $\{X_i\}_{i=1}^n$ .

- $E[|Y_i|] = E[|E[f(X)|X_1, \dots, X_i]|] \leq E[|f(X)|] < \infty$ . (Use Jensen on  $|(\cdot)|$ )
- Furthermore,

$$\begin{aligned} E[Y_{i+1}|X_1, \dots, X_i] &= E[E[f(X)|X_1, \dots, X_{i+1}]|X_1, \dots, X_i] \\ &= E[f(X)|X_1, \dots, X_i] = Y_i \quad \text{The tower property} \end{aligned}$$

# Likelihood ratio

## Example

Let  $f, g$  be two densities such that  $g$  is absolutely continuous w.r.t  $f$ .

Suppose  $\{X_i\}_{i=1}^{\infty} \stackrel{iid}{\sim} f$  and  $Y_n$  is the likelihood ratio  $\prod_{i=1}^n \frac{g(X_i)}{f(X_i)}$  for the first  $n$  datapoints. Then  $\{Y_n\}$  forms a Martingale sequence w.r.t  $\{X_n\}$ .

# Likelihood ratio

## Example

Let  $f, g$  be two densities such that  $g$  is absolutely continuous w.r.t  $f$ .

Suppose  $\{X_i\}_{i=1}^{\infty} \stackrel{iid}{\sim} f$  and  $Y_n$  is the likelihood ratio  $\prod_{i=1}^n \frac{g(X_i)}{f(X_i)}$  for the first  $n$  datapoints. Then  $\{Y_n\}$  forms a Martingale sequence w.r.t  $\{X_n\}$ .

- First recall that  $E[|Y_n|] = E[Y_n] = 1$

# Likelihood ratio

## Example

Let  $f, g$  be two densities such that  $g$  is absolutely continuous w.r.t  $f$ .

Suppose  $\{X_i\}_{i=1}^{\infty} \stackrel{iid}{\sim} f$  and  $Y_n$  is the likelihood ratio  $\prod_{i=1}^n \frac{g(X_i)}{f(X_i)}$  for the first  $n$  datapoints. Then  $\{Y_n\}$  forms a Martingale sequence w.r.t  $\{X_n\}$ .

- First recall that  $E[|Y_n|] = E[Y_n] = 1$
- 

$$\begin{aligned} E[Y_{n+1}|X_1, \dots, X_n] &= E \left[ \prod_{i=1}^{n+1} \frac{g(X_i)}{f(X_i)} \middle| X_1, \dots, X_n \right] \\ &= \prod_{i=1}^n \frac{g(X_i)}{f(X_i)} E \left[ \frac{g(X_{n+1})}{f(X_{n+1})} \right] = Y_n \end{aligned}$$

# Martingale Difference Sequence

## Definition

A sequence  $\{D_i\}$  of random variables adapted to a filtration  $\{\mathcal{F}_i\}$  is a Martingale Difference Sequence if,

$$E[|D_i|] < \infty \quad E[D_{i+1}|\mathcal{F}_i] = 0$$

# Martingale Difference Sequence

## Definition

A sequence  $\{D_i\}$  of random variables adapted to a filtration  $\{\mathcal{F}_i\}$  is a Martingale Difference Sequence if,

$$E[|D_i|] < \infty \quad E[D_{i+1}|\mathcal{F}_i] = 0$$

- Let  $\{Y_i\}$  be a martingale sequence.

# Martingale Difference Sequence

## Definition

A sequence  $\{D_i\}$  of random variables adapted to a filtration  $\{\mathcal{F}_i\}$  is a Martingale Difference Sequence if,

$$E[|D_i|] < \infty \quad E[D_{i+1}|\mathcal{F}_i] = 0$$

- Let  $\{Y_i\}$  be a martingale sequence.
- Then  $D_{i+1} = Y_{i+1} - Y_i$  define a Martingale Difference Sequence.
- $E[D_{i+1}|\mathcal{F}_i] = E[Y_{i+1}|\mathcal{F}_i] - E[Y_i|\mathcal{F}_i] = Y_i - Y_i = 0$ .

# Martingale Difference Sequence

## Definition

A sequence  $\{D_i\}$  of random variables adapted to a filtration  $\{\mathcal{F}_i\}$  is a Martingale Difference Sequence if,

$$E[|D_i|] < \infty \quad E[D_{i+1}|\mathcal{F}_i] = 0$$

- Let  $\{Y_i\}$  be a martingale sequence.
- Then  $D_{i+1} = Y_{i+1} - Y_i$  define a Martingale Difference Sequence.
- $E[D_{i+1}|\mathcal{F}_i] = E[Y_{i+1}|\mathcal{F}_i] - E[Y_i|\mathcal{F}_i] = Y_i - Y_i = 0$ .
  - $E[Y_{i+1}|\mathcal{F}_i] = Y_i$  because of the martingale property,
  - $E[Y_i|\mathcal{F}_i] = Y_i$  since  $Y_i$  is measurable w.r.t the filtration  $\mathcal{F}_i$ .



# Concentration inequalities

## Theorem

Consider a Martingale sequence  $\{D_i\}$  (adapted to a filtration  $\{\mathcal{F}_i\}$ ) that satisfies  $E[e^{\lambda D_i} | \mathcal{F}_{i-1}] \leq e^{\lambda^2 \nu_i^2 / 2}$  a.s. for any  $|\lambda| < 1/b_i$ .

- The sum  $\sum_i D_i$  is sub-exponential with parameters  $(\sqrt{\sum_k \nu_k^2}, b_*)$  where  $b_* := \max_i b_i$ .
- Hence for all  $t \geq 0$ ,

$$P \left[ \left| \sum_{i=1}^n D_i \right| \geq t \right] \leq \begin{cases} 2e^{-\frac{t^2}{2 \sum_k \nu_k^2}} & \text{If } 0 \leq t \leq \frac{\sum_k \nu_k^2}{b_*} \\ 2e^{-\frac{t}{2b_*}} & \text{If } t > \frac{\sum_k \nu_k^2}{b_*} \end{cases}$$

## Proof.

$$\text{Let } X := \sum_{i=1}^n D_i.$$

$$\begin{aligned} E[e^{\lambda \sum_i D_i}] &= E[E[e^{\lambda \sum_i D_i} | \mathcal{F}_{n-1}]] = E[e^{\lambda \sum_{i=1}^{n-1} D_i} E[e^{\lambda D_n} | \mathcal{F}_{n-1}]] \\ &\leq E[e^{\lambda \sum_{i=1}^{n-1} D_i}] e^{\lambda^2 \nu_n^2 / 2} \quad \text{If } |\lambda| < 1/b_n \\ &\leq E[e^{\lambda \sum_{i=1}^{n-2} D_i}] e^{\lambda^2 (\nu_{n-1}^2 + \nu_n^2) / 2} \quad \text{If } |\lambda| < 1/b_n, 1/b_{n-1} \\ &\leq e^{\sum_i \lambda^2 \nu_i^2 / 2} \quad \text{If } |\lambda| < \min_i 1/b_i \end{aligned}$$

Using our previous theorem on sub-exponential random variables, the result is proven in one direction. The other direction is identical leading to the factor of 2.  $\square$

# Azuma-Hoeffding

## Corollary (Azuma-Hoeffding)

Let  $\{D_k\}$  be a Martingale Difference Sequence adapted to the filtration  $\{\mathcal{F}_k\}$  and suppose  $|D_k| \leq b_k$  a.s. for all  $k \geq 1$ . Then  $\forall t \geq 0$ ,

$$P \left[ \left| \sum_{k=1}^n D_k \right| \geq t \right] \leq 2e^{-\frac{t^2}{2 \sum_k b_k^2}}$$

## Proof.

- We can rework the last proof. We need  $|E[e^{\lambda D_n} | \mathcal{F}_{n-1}]|$ .
- This is bounded by  $e^{\lambda^2 b_n^2 / 2}$ , since  $D_n$  is mean zero sub-gaussian with  $\sigma = b_n$ .



# McDiarmid's inequality

## Theorem

Let  $f : \mathcal{X}^n \rightarrow \mathbb{R}$  satisfy the following bounded difference condition

$\forall x_1, \dots, x_n, x'_i \in \mathcal{X}$ :

$$|f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq B_i,$$

then,  $P(|f(X) - E[f(X)]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_i B_i^2}\right)$

- Note that this boils down to Hoeffding's when  $f$  is the sum of bounded random variables.

## Proof.

- Define  $Y_i = E[f(X)|\mathcal{F}_i]$  and  $D_i = Y_i - Y_{i-1}$ .

## Proof.

- Define  $Y_i = E[f(X)|\mathcal{F}_i]$  and  $D_i = Y_i - Y_{i-1}$ .
- Since  $\{Y_i\}$  is a Martingale sequence w.r.t  $\{X_i\}$ ,  $\{D_i\}$  is a Martingale difference sequence.

## Proof.

- Define  $Y_i = E[f(X)|\mathcal{F}_i]$  and  $D_i = Y_i - Y_{i-1}$ .
- Since  $\{Y_i\}$  is a Martingale sequence w.r.t  $\{X_i\}$ ,  $\{D_i\}$  is a Martingale difference sequence.
- We have:

$$\begin{aligned} D_i &= E[f(X)|\mathcal{F}_i] - E[f(X)|\mathcal{F}_{i-1}] \\ &= E[f(X)|X_1, \dots, X_i] - E[f(X)|X_1, \dots, X_{i-1}] \\ &\leq \sup_x (E[f(X)|X_1, \dots, x] - E[f(X)|X_1, \dots, X_{i-1}]) =: U_i \\ D_i &\geq \inf_x (E[f(X)|X_1, \dots, x] - E[f(X)|X_1, \dots, X_{i-1}]) =: L_i \end{aligned}$$

## Proof.

- Define  $Y_i = E[f(X)|\mathcal{F}_i]$  and  $D_i = Y_i - Y_{i-1}$ .
- Since  $\{Y_i\}$  is a Martingale sequence w.r.t  $\{X_i\}$ ,  $\{D_i\}$  is a Martingale difference sequence.
- We have:

$$\begin{aligned} D_i &= E[f(X)|\mathcal{F}_i] - E[f(X)|\mathcal{F}_{i-1}] \\ &= E[f(X)|X_1, \dots, X_i] - E[f(X)|X_1, \dots, X_{i-1}] \\ &\leq \sup_x (E[f(X)|X_1, \dots, x] - E[f(X)|X_1, \dots, X_{i-1}]) =: U_i \\ D_i &\geq \inf_x (E[f(X)|X_1, \dots, x] - E[f(X)|X_1, \dots, X_{i-1}]) =: L_i \end{aligned}$$

- We also have:

$$U_i - L_i \leq B_i$$



## Proof.

- Define  $Y_i = E[f(X)|\mathcal{F}_i]$  and  $D_i = Y_i - Y_{i-1}$ .
- Since  $\{Y_i\}$  is a Martingale sequence w.r.t  $\{X_i\}$ ,  $\{D_i\}$  is a Martingale difference sequence.
- We have:

$$\begin{aligned} D_i &= E[f(X)|\mathcal{F}_i] - E[f(X)|\mathcal{F}_{i-1}] \\ &= E[f(X)|X_1, \dots, X_i] - E[f(X)|X_1, \dots, X_{i-1}] \\ &\leq \sup_x (E[f(X)|X_1, \dots, x] - E[f(X)|X_1, \dots, X_{i-1}]) =: U_i \\ D_i &\geq \inf_x (E[f(X)|X_1, \dots, x] - E[f(X)|X_1, \dots, X_{i-1}]) =: L_i \end{aligned}$$

- We also have:

$$U_i - L_i \leq B_i$$

- Now apply Azuma-Hoeffding.

## Example: Mean absolute deviation

### Example

Consider an i.i.d random variable sequence  $\{X_k\}_{k=1}^{\infty}$  with  $|X_k| \leq b$ . Define the mean absolute deviation:

$$U = \frac{1}{\binom{n}{2}} \sum_{j \neq k} |X_j - X_k|$$

As we will see later, the above is a type of a pairwise U-Statistics. We want to bound  $|U - E[U]|$ .

- Note that the summands are not independent.

## Example: Mean absolute deviation

### Example

Consider an i.i.d random variable sequence  $\{X_k\}_{k=1}^{\infty}$  with  $|X_k| \leq b$ . Define the mean absolute deviation:

$$U = \frac{1}{\binom{n}{2}} \sum_{j \neq k} |X_j - X_k|$$

As we will see later, the above is a type of a pairwise U-Statistics. We want to bound  $|U - E[U]|$ .

- Note that the summands are not independent.
- Also note that  $||X_i - X_j| - |X_i - X'_j|| \leq |X_j - X'_j| \leq 2b$

## Example: Mean absolute deviation

### Example

Consider an i.i.d random variable sequence  $\{X_k\}_{k=1}^{\infty}$  with  $|X_k| \leq b$ . Define the mean absolute deviation:

$$U = \frac{1}{\binom{n}{2}} \sum_{j \neq k} |X_j - X_k|$$

As we will see later, the above is a type of a pairwise U-Statistics. We want to bound  $|U - E[U]|$ .

- Note that the summands are not independent.
- Also note that  $||X_i - X_j| - |X_i - X'_j|| \leq |X_j - X'_j| \leq 2b$
- So  $|U(x_1, \dots, x_i, \dots, x_n) - U(x_1, \dots, x'_i, \dots, x_n)| \leq \frac{(n-1)2b}{\binom{n}{2}} = \frac{4b}{n}$

## Example: Mean absolute deviation

### Example

Consider an i.i.d random variable sequence  $\{X_k\}_{k=1}^{\infty}$  with  $|X_k| \leq b$ . Define the mean absolute deviation:

$$U = \frac{1}{\binom{n}{2}} \sum_{j \neq k} |X_j - X_k|$$

As we will see later, the above is a type of a pairwise U-Statistics. We want to bound  $|U - E[U]|$ .

- Note that the summands are not independent.
- Also note that  $||X_i - X_j| - |X_i - X'_j|| \leq |X_j - X'_j| \leq 2b$
- So  $|U(x_1, \dots, x_i, \dots, x_n) - U(x_1, \dots, x'_i, \dots, x_n)| \leq \frac{(n-1)2b}{\binom{n}{2}} = \frac{4b}{n}$
- Use McDiarmid's inequality,  $P(|U - E[U]| \geq t) \leq 2 \exp\left(\frac{-nt^2}{8b^2}\right)$

# Lipschitz functions of Gaussian random variables

## Definition

A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L$ -Lipschitz w.r.t the Euclidean norm if

$$|f(x) - f(y)| \leq L\|x - y\|_2 \quad \forall x, y \in \mathbb{R}^n$$

## Theorem

*Let  $(X_1, \dots, X_n)$  be a vector of iid  $N(0, 1)$  random variables. Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be  $L$ -Lipschitz w.r.t the Euclidean norm. Then  $f(X) - E[f(X)]$  is sub-gaussian with parameter at most  $L$ , i.e.  $\forall t \geq 0$ ,*

$$P(|f(X) - E[f(X)]| \geq t) \leq e^{-\frac{t^2}{2L}}$$

- A  $L$  Lipschitz function of a vector of i.i.d  $N(0, 1)$  random variables concentrate like a  $N(0, L^2)$  random variable, irrespective of how long the vector was.



