

## Lecture 2

July 1, 2025

## Detour: Martingales

- ▶ A filtration  $\{\mathcal{F}_k\}_{k \geq 1}$  is a nested sequence of  $\sigma$ -fields  
 $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$

## Detour: Martingales

- ▶ A filtration  $\{\mathcal{F}_k\}_{k \geq 1}$  is a nested sequence of  $\sigma$ -fields  
 $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$
- ▶ A sequence of random variables  $Y_1, Y_2, \dots$

## Detour: Martingales

- ▶ A filtration  $\{\mathcal{F}_k\}_{k \geq 1}$  is a nested sequence of  $\sigma$ -fields  
 $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$
- ▶ A sequence of random variables  $Y_1, Y_2, \dots$
- ▶ The pair  $\{(Y_k, \mathcal{F}_k)\}$  is a martingale if

## Detour: Martingales

- ▶ A filtration  $\{\mathcal{F}_k\}_{k \geq 1}$  is a nested sequence of  $\sigma$ -fields  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$
- ▶ A sequence of random variables  $Y_1, Y_2, \dots$
- ▶ The pair  $\{(Y_k, \mathcal{F}_k)\}$  is a martingale if
  - ▶  $\{Y_k\}_{k \geq 1}$  is adapted  $\{\mathcal{F}_k\}_{k \geq 1}$ , i.e.,  $Y_k \in \mathcal{F}_k$  for all  $k \geq 1$ .

## Detour: Martingales

- ▶ A filtration  $\{\mathcal{F}_k\}_{k \geq 1}$  is a nested sequence of  $\sigma$ -fields  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$
- ▶ A sequence of random variables  $Y_1, Y_2, \dots$
- ▶ The pair  $\{(Y_k, \mathcal{F}_k)\}$  is a martingale if
  - ▶  $\{Y_k\}_{k \geq 1}$  is adapted  $\{\mathcal{F}_k\}_{k \geq 1}$ , i.e.,  $Y_k \in \mathcal{F}_k$  for all  $k \geq 1$ .
  - ▶  $\mathbb{E}(|Y_k|) < \infty$ .

## Detour: Martingales

- ▶ A filtration  $\{\mathcal{F}_k\}_{k \geq 1}$  is a nested sequence of  $\sigma$ -fields  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$
- ▶ A sequence of random variables  $Y_1, Y_2, \dots$
- ▶ The pair  $\{(Y_k, \mathcal{F}_k)\}$  is a martingale if
  - ▶  $\{Y_k\}_{k \geq 1}$  is adapted  $\{\mathcal{F}_k\}_{k \geq 1}$ , i.e.,  $Y_k \in \mathcal{F}_k$  for all  $k \geq 1$ .
  - ▶  $\mathbb{E}(|Y_k|) < \infty$ .
  - ▶  $\mathbb{E}(Y_{k+1} | \mathcal{F}_k) = Y_k$  for all  $k \geq 1$ .

## Detour: Martingales

- ▶ A filtration  $\{\mathcal{F}_k\}_{k \geq 1}$  is a nested sequence of  $\sigma$ -fields  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$
- ▶ A sequence of random variables  $Y_1, Y_2, \dots$
- ▶ The pair  $\{(Y_k, \mathcal{F}_k)\}$  is a martingale if
  - ▶  $\{Y_k\}_{k \geq 1}$  is adapted  $\{\mathcal{F}_k\}_{k \geq 1}$ , i.e.,  $Y_k \in \mathcal{F}_k$  for all  $k \geq 1$ .
  - ▶  $\mathbb{E}(|Y_k|) < \infty$ .
  - ▶  $\mathbb{E}(Y_{k+1} | \mathcal{F}_k) = Y_k$  for all  $k \geq 1$ .
- ▶ Often  $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ , in which case we say  $\{Y_k\}$  is martingale with respect to  $\{X_k\}$ .



## Detour: Martingales

- ▶ A filtration  $\{\mathcal{F}_k\}_{k \geq 1}$  is a nested sequence of  $\sigma$ -fields  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$
- ▶ A sequence of random variables  $Y_1, Y_2, \dots$
- ▶ The pair  $\{(Y_k, \mathcal{F}_k)\}$  is a martingale if
  - ▶  $\{Y_k\}_{k \geq 1}$  is adapted  $\{\mathcal{F}_k\}_{k \geq 1}$ , i.e.,  $Y_k \in \mathcal{F}_k$  for all  $k \geq 1$ .
  - ▶  $\mathbb{E}(|Y_k|) < \infty$ .
  - ▶  $\mathbb{E}(Y_{k+1}|\mathcal{F}_k) = Y_k$  for all  $k \geq 1$ .
- ▶ Often  $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ , in which case we say  $\{Y_k\}$  is martingale with respect to  $\{X_k\}$ . The key condition in this case is

$$\mathbb{E}(Y_{k+1}|X_1, \dots, X_k) = Y_k$$

## Detour: Martingales

- ▶ A filtration  $\{\mathcal{F}_k\}_{k \geq 1}$  is a nested sequence of  $\sigma$ -fields  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$
- ▶ A sequence of random variables  $Y_1, Y_2, \dots$
- ▶ The pair  $\{(Y_k, \mathcal{F}_k)\}$  is a martingale if
  - ▶  $\{Y_k\}_{k \geq 1}$  is adapted  $\{\mathcal{F}_k\}_{k \geq 1}$ , i.e.,  $Y_k \in \mathcal{F}_k$  for all  $k \geq 1$ .
  - ▶  $\mathbb{E}(|Y_k|) < \infty$ .
  - ▶  $\mathbb{E}(Y_{k+1}|\mathcal{F}_k) = Y_k$  for all  $k \geq 1$ .
- ▶ Often  $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ , in which case we say  $\{Y_k\}$  is martingale with respect to  $\{X_k\}$ . The key condition in this case is

$$\mathbb{E}(Y_{k+1}|X_1, \dots, X_k) = Y_k$$

- ▶ One of the most general dependence structures in probability.

## Detour: Martingales

- ▶ A filtration  $\{\mathcal{F}_k\}_{k \geq 1}$  is a nested sequence of  $\sigma$ -fields  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$
- ▶ A sequence of random variables  $Y_1, Y_2, \dots$
- ▶ The pair  $\{(Y_k, \mathcal{F}_k)\}$  is a martingale if
  - ▶  $\{Y_k\}_{k \geq 1}$  is adapted  $\{\mathcal{F}_k\}_{k \geq 1}$ , i.e.,  $Y_k \in \mathcal{F}_k$  for all  $k \geq 1$ .
  - ▶  $\mathbb{E}(|Y_k|) < \infty$ .
  - ▶  $\mathbb{E}(Y_{k+1}|\mathcal{F}_k) = Y_k$  for all  $k \geq 1$ .
- ▶ Often  $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ , in which case we say  $\{Y_k\}$  is martingale with respect to  $\{X_k\}$ . The key condition in this case is

$$\mathbb{E}(Y_{k+1}|X_1, \dots, X_k) = Y_k$$

- ▶ One of the most general dependence structures in probability.
- ▶ Allow relaxing independence assumptions in classical limit theorems.

## Detour: Martingale Examples

- ▶ Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ .

## Detour: Martingale Examples

- ▶ Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ . Simply notice that

$$\mathbb{E}(S_{k+1} | X_1, \dots, X_k) = \mathbb{E}\left(\sum_{i=1}^{k+1} X_i \mid X_1, \dots, X_k\right)$$

## Detour: Martingale Examples

- ▶ Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ . Simply notice that

$$\begin{aligned}\mathbb{E}(S_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\sum_{i=1}^{k+1} X_i \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}(\sum_{i=1}^k X_i + X_{k+1} | X_1, \dots, X_k)\end{aligned}$$

## Detour: Martingale Examples

- ▶ Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ . Simply notice that

$$\begin{aligned}\mathbb{E}(S_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\sum_{i=1}^{k+1} X_i \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}\left(\sum_{i=1}^k X_i + X_{k+1} \mid X_1, \dots, X_k\right) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1} | X_1, \dots, X_k)\end{aligned}$$

## Detour: Martingale Examples

- ▶ Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ . Simply notice that

$$\begin{aligned}\mathbb{E}(S_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\sum_{i=1}^{k+1} X_i \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}\left(\sum_{i=1}^k X_i + X_{k+1} \mid X_1, \dots, X_k\right) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1})\end{aligned}$$



## Detour: Martingale Examples

- ▶ Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ . Simply notice that

$$\begin{aligned}\mathbb{E}(S_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\sum_{i=1}^{k+1} X_i \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}(\sum_{i=1}^k X_i + X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1}) \\ &= \sum_{i=1}^k X_i\end{aligned}$$

## Detour: Martingale Examples

- ▶ Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ . Simply notice that

$$\begin{aligned}\mathbb{E}(S_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\sum_{i=1}^{k+1} X_i \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}\left(\sum_{i=1}^k X_i + X_{k+1} \mid X_1, \dots, X_k\right) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1}) \\ &= \sum_{i=1}^k X_i = S_k\end{aligned}$$

## Detour: Martingale Examples

- ▶ Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ . Simply notice that

$$\begin{aligned}\mathbb{E}(S_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\sum_{i=1}^{k+1} X_i \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}(\sum_{i=1}^k X_i + X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1}) \\ &= \sum_{i=1}^k X_i = S_k\end{aligned}$$

- ▶ Partial products  $L_k = \prod_{j=1}^k X_j$  of an iid sequence with  $\mathbb{E}(X_i) = 1$ .

## Detour: Martingale Examples

- ▶ Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ . Simply notice that

$$\begin{aligned}\mathbb{E}(S_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\sum_{i=1}^{k+1} X_i \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}(\sum_{i=1}^k X_i + X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1}) \\ &= \sum_{i=1}^k X_i = S_k\end{aligned}$$

- ▶ Partial products  $L_k = \prod_{j=1}^k X_j$  of an iid sequence with  $\mathbb{E}(X_i) = 1$ .

$$\mathbb{E}(L_{k+1} | X_1, \dots, X_k) = \mathbb{E}\left(\prod_{j=1}^{k+1} X_j \mid X_1, \dots, X_k\right)$$

## Detour: Martingale Examples

- ▶ Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ . Simply notice that

$$\begin{aligned}\mathbb{E}(S_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\sum_{i=1}^{k+1} X_i \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}(\sum_{i=1}^k X_i + X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1}) \\ &= \sum_{i=1}^k X_i = S_k\end{aligned}$$

- ▶ Partial products  $L_k = \prod_{j=1}^k X_j$  of an iid sequence with  $\mathbb{E}(X_i) = 1$ .

$$\begin{aligned}\mathbb{E}(L_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\prod_{j=1}^{k+1} X_j \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \cdot \prod_{j=1}^k X_j\end{aligned}$$

## Detour: Martingale Examples

- Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ . Simply notice that

$$\begin{aligned}\mathbb{E}(S_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\sum_{i=1}^{k+1} X_i \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}(\sum_{i=1}^k X_i + X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1}) \\ &= \sum_{i=1}^k X_i = S_k\end{aligned}$$

- Partial products  $L_k = \prod_{j=1}^k X_j$  of an iid sequence with  $\mathbb{E}(X_i) = 1$ .

$$\begin{aligned}\mathbb{E}(L_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\prod_{j=1}^{k+1} X_j \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \cdot \prod_{j=1}^k X_j \\ &= \mathbb{E}(X_{k+1}) \cdot \prod_{j=1}^k X_j\end{aligned}$$

## Detour: Martingale Examples

- Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ . Simply notice that

$$\begin{aligned}\mathbb{E}(S_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\sum_{i=1}^{k+1} X_i \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}(\sum_{i=1}^k X_i + X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1}) \\ &= \sum_{i=1}^k X_i = S_k\end{aligned}$$

- Partial products  $L_k = \prod_{j=1}^k X_j$  of an iid sequence with  $\mathbb{E}(X_i) = 1$ .

$$\begin{aligned}\mathbb{E}(L_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\prod_{j=1}^{k+1} X_j \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \cdot \prod_{j=1}^k X_j \\ &= \mathbb{E}(X_{k+1}) \cdot \prod_{j=1}^k X_j \\ &= 1 \cdot \prod_{j=1}^k X_j.\end{aligned}$$

## Detour: Martingale Examples

- Partial sums  $S_k = \sum_{i=1}^k X_i$  of an iid zero-mean sequence  $\{X_k\}$ . Simply notice that

$$\begin{aligned}\mathbb{E}(S_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\sum_{i=1}^{k+1} X_i \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}(\sum_{i=1}^k X_i + X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \\ &= \sum_{i=1}^k X_i + \mathbb{E}(X_{k+1}) \\ &= \sum_{i=1}^k X_i = S_k\end{aligned}$$

- Partial products  $L_k = \prod_{j=1}^k X_j$  of an iid sequence with  $\mathbb{E}(X_i) = 1$ .

$$\begin{aligned}\mathbb{E}(L_{k+1} | X_1, \dots, X_k) &= \mathbb{E}\left(\prod_{j=1}^{k+1} X_j \mid X_1, \dots, X_k\right) \\ &= \mathbb{E}(X_{k+1} | X_1, \dots, X_k) \cdot \prod_{j=1}^k X_j \\ &= \mathbb{E}(X_{k+1}) \cdot \prod_{j=1}^k X_j \\ &= 1 \cdot \prod_{j=1}^k X_j. \\ &= \prod_{j=1}^k X_j.\end{aligned}$$



- ▶ Likelihood ratio process is an example. Say the  $X_i$  are i.i.d from the pdf  $f_0$ . Then for any pdf  $f_1$  we have that

$$L_k = \prod_{j=1}^k \frac{f_1(X_j)}{f_0(X_j)}.$$

- Likelihood ratio process is an example. Say the  $X_i$  are i.i.d from the pdf  $f_0$ . Then for any pdf  $f_1$  we have that

$$L_k = \prod_{j=1}^k \frac{f_1(X_j)}{f_0(X_j)}.$$

This is a martingale since

$$\mathbb{E} \left( \frac{f_1(X_i)}{f_0(X_i)} \right) = \int \frac{f_1(x)}{f_0(x)} f_0(x) dx = \int f_1(x) dx = 1.$$

- ▶ Likelihood ratio process is an example. Say the  $X_i$  are i.i.d from the pdf  $f_0$ . Then for any pdf  $f_1$  we have that

$$L_k = \prod_{j=1}^k \frac{f_1(X_j)}{f_0(X_j)}.$$

This is a martingale since

$$\mathbb{E} \left( \frac{f_1(X_i)}{f_0(X_i)} \right) = \int \frac{f_1(x)}{f_0(x)} f_0(x) dx = \int f_1(x) dx = 1.$$

- ▶ Doob's martingale: For any integrable  $Z$  (i.e.  $\mathbb{E}(|Z|) < \infty$ ) the following is always a martingale:

$$Y_k := \mathbb{E}(Z | X_1, \dots, X_k).$$

- ▶ Likelihood ratio process is an example. Say the  $X_i$  are i.i.d from the pdf  $f_0$ . Then for any pdf  $f_1$  we have that

$$L_k = \prod_{j=1}^k \frac{f_1(X_j)}{f_0(X_j)}.$$

This is a martingale since

$$\mathbb{E} \left( \frac{f_1(X_i)}{f_0(X_i)} \right) = \int \frac{f_1(x)}{f_0(x)} f_0(x) dx = \int f_1(x) dx = 1.$$

- ▶ Doob's martingale: For any integrable  $Z$  (i.e.  $\mathbb{E}(|Z|) < \infty$ ) the following is always a martingale:

$$Y_k := \mathbb{E}(Z | X_1, \dots, X_k).$$

Notice that

$$\mathbb{E}(Y_{k+1} | X_1, \dots, X_k) = \mathbb{E}(\mathbb{E}(Z | X_1, \dots, X_{k+1}) | X_1, \dots, X_k)$$

- Likelihood ratio process is an example. Say the  $X_i$  are i.i.d from the pdf  $f_0$ . Then for any pdf  $f_1$  we have that

$$L_k = \prod_{j=1}^k \frac{f_1(X_j)}{f_0(X_j)}.$$

This is a martingale since

$$\mathbb{E} \left( \frac{f_1(X_i)}{f_0(X_i)} \right) = \int \frac{f_1(x)}{f_0(x)} f_0(x) dx = \int f_1(x) dx = 1.$$

- Doob's martingale: For any integrable  $Z$  (i.e.  $\mathbb{E}(|Z|) < \infty$ ) the following is always a martingale:

$$Y_k := \mathbb{E}(Z | X_1, \dots, X_k).$$

Notice that

$$\begin{aligned} \mathbb{E}(Y_{k+1} | X_1, \dots, X_k) &= \mathbb{E}(\mathbb{E}(Z | X_1, \dots, X_{k+1}) | X_1, \dots, X_k) \\ &= \mathbb{E}(Z | X_1, \dots, X_k) \end{aligned}$$

- Likelihood ratio process is an example. Say the  $X_i$  are i.i.d from the pdf  $f_0$ . Then for any pdf  $f_1$  we have that

$$L_k = \prod_{j=1}^k \frac{f_1(X_j)}{f_0(X_j)}.$$

This is a martingale since

$$\mathbb{E} \left( \frac{f_1(X_i)}{f_0(X_i)} \right) = \int \frac{f_1(x)}{f_0(x)} f_0(x) dx = \int f_1(x) dx = 1.$$

- Doob's martingale: For any integrable  $Z$  (i.e.  $\mathbb{E}(|Z|) < \infty$ ) the following is always a martingale:

$$Y_k := \mathbb{E}(Z | X_1, \dots, X_k).$$

Notice that

$$\begin{aligned} \mathbb{E}(Y_{k+1} | X_1, \dots, X_k) &= \mathbb{E}(\mathbb{E}(Z | X_1, \dots, X_{k+1}) | X_1, \dots, X_k) \\ &= \mathbb{E}(Z | X_1, \dots, X_k) \\ &= Y_k. \end{aligned}$$

- **Theorem 7 (Azuma–Hoeffding).** Let  $X = (X_1, \dots, X_n)^\top$  be a random vector and let  $Z = f(X)$ . Consider the Doob's martingale

$$Y_i := \mathbb{E}_i(Z) := \mathbb{E}(Z | X_1, \dots, X_i)$$

and let  $\Delta_i = Y_i - Y_{i-1}$ . Assume that

$$\mathbb{E}_{i-1} \left( e^{\lambda \Delta_i} \right) \leq e^{\sigma_i^2 \lambda^2 / 2}, \quad \forall \lambda \in \mathbb{R} \quad (1)$$

almost surely for all  $i = 1, \dots, n$ . Then,  $Z - \mathbb{E}(Z)$  is sub-Gaussian with parameter  $\sigma = \sqrt{\sum_{i=1}^n \sigma_i^2}$ .

- **Theorem 7 (Azuma–Hoeffding).** Let  $X = (X_1, \dots, X_n)^\top$  be a random vector and let  $Z = f(X)$ . Consider the Doob's martingale

$$Y_i := \mathbb{E}_i(Z) := \mathbb{E}(Z | X_1, \dots, X_i)$$

and let  $\Delta_i = Y_i - Y_{i-1}$ . Assume that

$$\mathbb{E}_{i-1} \left( e^{\lambda \Delta_i} \right) \leq e^{\sigma_i^2 \lambda^2 / 2}, \quad \forall \lambda \in \mathbb{R} \quad (1)$$

almost surely for all  $i = 1, \dots, n$ . Then,  $Z - \mathbb{E}(Z)$  is sub-Gaussian with parameter  $\sigma = \sqrt{\sum_{i=1}^n \sigma_i^2}$ .

- In particular, we have the tail bound

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq t) \leq 2 \exp \left( -\frac{t^2}{2\sigma^2} \right).$$



## Proof

- ▶ Let  $S_j = \sum_{i=1}^j \Delta_i$  which is only a function of  $X_i$  for  $i \leq j$ .

## Proof

- ▶ Let  $S_j = \sum_{i=1}^j \Delta_i$  which is only a function of  $X_i$  for  $i \leq j$ .
- ▶ Noting that  $\mathbb{E}_0(Z) = \mathbb{E}(Z)$  and  
 $Y_n = \mathbb{E}_n(Z) = \mathbb{E}(f(X)|X_1, \dots, X_n) = f(X) = Z,$

$$S_n = \sum_{i=1}^n \Delta_i = \sum_{i=1}^n (Y_i - Y_{i-1}) = Y_n - Y_0 = Z - \mathbb{E}(Z).$$

## Proof

- ▶ Let  $S_j = \sum_{i=1}^j \Delta_i$  which is only a function of  $X_i$  for  $i \leq j$ .
- ▶ Noting that  $\mathbb{E}_0(Z) = \mathbb{E}(Z)$  and  
 $Y_n = \mathbb{E}_n(Z) = \mathbb{E}(f(X)|X_1, \dots, X_n) = f(X) = Z,$

$$S_n = \sum_{i=1}^n \Delta_i = \sum_{i=1}^n (Y_i - Y_{i-1}) = Y_n - Y_0 = Z - \mathbb{E}(Z).$$

- ▶ By properties of conditional expectation, and assumption (1)

$$\mathbb{E}_{n-1}(e^{\lambda S_n}) = e^{\lambda S_{n-1}} \mathbb{E}_{n-1} \left( e^{\lambda \Delta_n} \right)$$

## Proof

- ▶ Let  $S_j = \sum_{i=1}^j \Delta_i$  which is only a function of  $X_i$  for  $i \leq j$ .
- ▶ Noting that  $\mathbb{E}_0(Z) = \mathbb{E}(Z)$  and  
 $Y_n = \mathbb{E}_n(Z) = \mathbb{E}(f(X)|X_1, \dots, X_n) = f(X) = Z,$

$$S_n = \sum_{i=1}^n \Delta_i = \sum_{i=1}^n (Y_i - Y_{i-1}) = Y_n - Y_0 = Z - \mathbb{E}(Z).$$

- ▶ By properties of conditional expectation, and assumption (1)

$$\mathbb{E}_{n-1}(e^{\lambda S_n}) = e^{\lambda S_{n-1}} \mathbb{E}_{n-1} \left( e^{\lambda \Delta_n} \right) \leq e^{\lambda S_{n-1}} e^{\sigma_n^2 \lambda^2 / 2}.$$

## Proof

- ▶ Let  $S_j = \sum_{i=1}^j \Delta_i$  which is only a function of  $X_i$  for  $i \leq j$ .
- ▶ Noting that  $\mathbb{E}_0(Z) = \mathbb{E}(Z)$  and  
 $Y_n = \mathbb{E}_n(Z) = \mathbb{E}(f(X)|X_1, \dots, X_n) = f(X) = Z,$

$$S_n = \sum_{i=1}^n \Delta_i = \sum_{i=1}^n (Y_i - Y_{i-1}) = Y_n - Y_0 = Z - \mathbb{E}(Z).$$

- ▶ By properties of conditional expectation, and assumption (1)

$$\mathbb{E}_{n-1}(e^{\lambda S_n}) = e^{\lambda S_{n-1}} \mathbb{E}_{n-1} \left( e^{\lambda \Delta_n} \right) \leq e^{\lambda S_{n-1}} e^{\sigma_n^2 \lambda^2 / 2}.$$

- ▶ Taking  $\mathbb{E}_{n-2}$  of both sides:

$$\mathbb{E}_{n-2} \left( e^{\lambda S_n} \right) \leq$$

## Proof

- ▶ Let  $S_j = \sum_{i=1}^j \Delta_i$  which is only a function of  $X_i$  for  $i \leq j$ .
- ▶ Noting that  $\mathbb{E}_0(Z) = \mathbb{E}(Z)$  and  
 $Y_n = \mathbb{E}_n(Z) = \mathbb{E}(f(X)|X_1, \dots, X_n) = f(X) = Z,$

$$S_n = \sum_{i=1}^n \Delta_i = \sum_{i=1}^n (Y_i - Y_{i-1}) = Y_n - Y_0 = Z - \mathbb{E}(Z).$$

- ▶ By properties of conditional expectation, and assumption (1)

$$\mathbb{E}_{n-1}(e^{\lambda S_n}) = e^{\lambda S_{n-1}} \mathbb{E}_{n-1} \left( e^{\lambda \Delta_n} \right) \leq e^{\lambda S_{n-1}} e^{\sigma_n^2 \lambda^2 / 2}.$$

- ▶ Taking  $\mathbb{E}_{n-2}$  of both sides:

$$\mathbb{E}_{n-2} \left( e^{\lambda S_n} \right) \leq e^{\sigma_n^2 \lambda^2 / 2} \mathbb{E}_{n-2} (e^{\lambda S_{n-1}})$$

## Proof

- ▶ Let  $S_j = \sum_{i=1}^j \Delta_i$  which is only a function of  $X_i$  for  $i \leq j$ .
- ▶ Noting that  $\mathbb{E}_0(Z) = \mathbb{E}(Z)$  and  
 $Y_n = \mathbb{E}_n(Z) = \mathbb{E}(f(X)|X_1, \dots, X_n) = f(X) = Z,$

$$S_n = \sum_{i=1}^n \Delta_i = \sum_{i=1}^n (Y_i - Y_{i-1}) = Y_n - Y_0 = Z - \mathbb{E}(Z).$$

- ▶ By properties of conditional expectation, and assumption (1)

$$\mathbb{E}_{n-1}(e^{\lambda S_n}) = e^{\lambda S_{n-1}} \mathbb{E}_{n-1} \left( e^{\lambda \Delta_n} \right) \leq e^{\lambda S_{n-1}} e^{\sigma_n^2 \lambda^2 / 2}.$$

- ▶ Taking  $\mathbb{E}_{n-2}$  of both sides:

$$\mathbb{E}_{n-2} \left( e^{\lambda S_n} \right) \leq e^{\sigma_n^2 \lambda^2 / 2} \mathbb{E}_{n-2} (e^{\lambda S_{n-1}}) \leq e^{\lambda S_{n-2}} e^{(\sigma_n^2 + \sigma_{n-1}^2) \lambda^2 / 2}.$$

► Taking  $\mathbb{E}_{n-2}$  of both sides:

$$\mathbb{E}_{n-2} \left( e^{\lambda S_n} \right) \leq e^{\sigma_n^2 \lambda^2 / 2} \mathbb{E}_{n-2} (e^{\lambda S_{n-1}}) \leq e^{\lambda S_{n-2}} e^{(\sigma_n^2 + \sigma_{n-1}^2) \lambda^2 / 2}$$



- Taking  $\mathbb{E}_{n-2}$  of both sides:

$$\mathbb{E}_{n-2} \left( e^{\lambda S_n} \right) \leq e^{\sigma_n^2 \lambda^2 / 2} \mathbb{E}_{n-2} (e^{\lambda S_{n-1}}) \leq e^{\lambda S_{n-2}} e^{(\sigma_n^2 + \sigma_{n-1}^2) \lambda^2 / 2}$$

- Repeating the process, we get

$$\mathbb{E}_0 \left( e^{\lambda S_n} \right) \leq \exp \left( \frac{\lambda}{2} \sum_{i=1}^n \sigma_i^2 \right) = \exp \left( \frac{\lambda^2 \sigma^2}{2} \right).$$

- Taking  $\mathbb{E}_{n-2}$  of both sides:

$$\mathbb{E}_{n-2} \left( e^{\lambda S_n} \right) \leq e^{\sigma_n^2 \lambda^2 / 2} \mathbb{E}_{n-2} (e^{\lambda S_{n-1}}) \leq e^{\lambda S_{n-2}} e^{(\sigma_n^2 + \sigma_{n-1}^2) \lambda^2 / 2}$$

- Repeating the process, we get

$$\mathbb{E}_0 \left( e^{\lambda S_n} \right) \leq \exp \left( \frac{\lambda}{2} \sum_{i=1}^n \sigma_i^2 \right) = \exp \left( \frac{\lambda^2 \sigma^2}{2} \right).$$

- Or

$$\mathbb{E} \left( e^{\lambda(Z - \mathbb{E}(Z))} \right) \leq \exp \left( \frac{\lambda}{2} \sum_{i=1}^n \sigma_i^2 \right) = \exp \left( \frac{\lambda^2 \sigma^2}{2} \right).$$

and the claim follows.

## Bounded difference inequality

- ▶ Conditional sub-G. assump. holds under **bounded difference** property:

$$\begin{aligned} & \left| f(x_1, \dots, x_{i-1}, \mathbf{x}_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, \mathbf{x}'_i, x_{i+1}, \dots, x_n) \right| \\ & \leq L_i \end{aligned} \tag{2}$$

for all  $x_1, \dots, x_n, x'_i \in \mathcal{X}$  and some constants  $(L_1, \dots, L_n)$ .

## Bounded difference inequality

- ▶ Conditional sub-G. assump. holds under **bounded difference** property:

$$\begin{aligned} & \left| f(x_1, \dots, x_{i-1}, \mathbf{x}_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, \mathbf{x}'_i, x_{i+1}, \dots, x_n) \right| \\ & \leq L_i \end{aligned} \tag{2}$$

for all  $x_1, \dots, x_n, \mathbf{x}'_i \in \mathcal{X}$  and some constants  $(L_1, \dots, L_n)$ .

- ▶ **Theorem 8 (Bounded difference).** Assume that  $X = (X_1, \dots, X_n)^\top$  has independent coordinates, and assume that  $f : \mathcal{X}^n \rightarrow \mathbb{R}$  satisfies the **bounded difference** property (2). Then

$$\mathbb{P}(|f(X) - \mathbb{E}(f(X))| \geq t) \leq 2 \exp \left( - \frac{2t^2}{\sum_{i=1}^n L_i^2} \right), \quad t \geq 0.$$

## Bounded difference inequality

- ▶ Conditional sub-G. assump. holds under **bounded difference** property:

$$\begin{aligned} & \left| f(x_1, \dots, x_{i-1}, \mathbf{x}_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, \mathbf{x}'_i, x_{i+1}, \dots, x_n) \right| \\ & \leq L_i \end{aligned} \tag{2}$$

for all  $x_1, \dots, x_n, \mathbf{x}'_i \in \mathcal{X}$  and some constants  $(L_1, \dots, L_n)$ .

- ▶ **Theorem 8 (Bounded difference).** Assume that  $X = (X_1, \dots, X_n)^\top$  has independent coordinates, and assume that  $f : \mathcal{X}^n \rightarrow \mathbb{R}$  satisfies the **bounded difference** property (2). Then

$$\mathbb{P}(|f(X) - \mathbb{E}(f(X))| \geq t) \leq 2 \exp \left( - \frac{2t^2}{\sum_{i=1}^n L_i^2} \right), \quad t \geq 0.$$

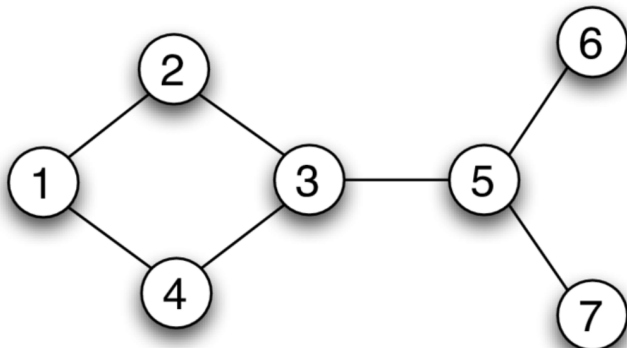
- ▶ **Example.**  $f(X) = \sum_{i=1}^n X_i$ ,  $X_i \in [a_i, b_i]$ .

# Clique number of Erdős-Rényi

- ▶ Let  $G$  be an undirected graph on  $n$  nodes.

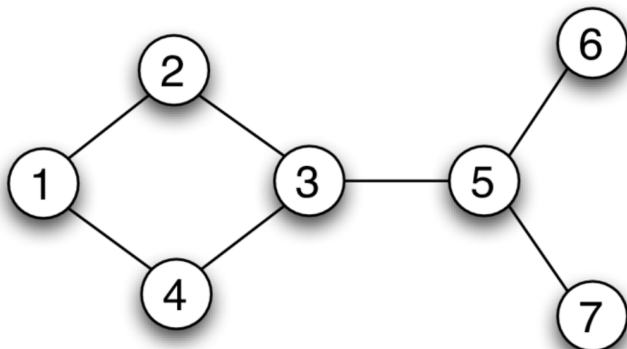
# Clique number of Erdős-Rényi

- Let  $G$  be an undirected graph on  $n$  nodes.



# Clique number of Erdős-Rényi

- ▶ Let  $G$  be an undirected graph on  $n$  nodes.



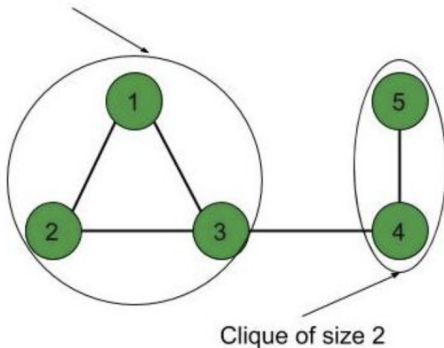
- ▶ The edge set is  
 $E(G) := \{\{1, 2\}, \{1, 4\}, \{2, 3\}, \{3, 4\}, \{3, 5\}, \{5, 6\}, \{5, 7\}\}.$



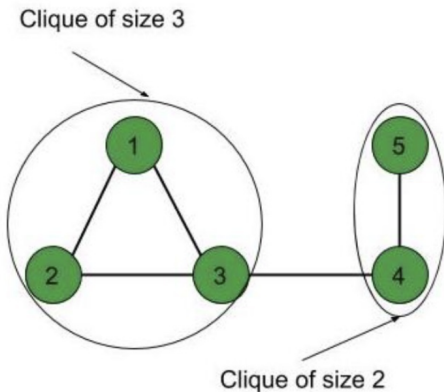
- ▶ A clique in  $G$  is a complete (induced) sub-graph.

- A clique in  $G$  is a complete (induced) sub-graph.

Clique of size 3



- ▶ A clique in  $G$  is a complete (induced) sub-graph.



- ▶ Clique number of  $G$ —denoted as  $\omega(G)$ —is the size of the largest clique(s).

- For two graphs  $G$  and  $G'$  that differ in at most 1 edge,

$$|\omega(G) - \omega(G')| \leq 1.$$

- ▶ For two graphs  $G$  and  $G'$  that differ in at most 1 edge,

$$|\omega(G) - \omega(G')| \leq 1.$$

- ▶ Thus  $E(G) \rightarrow \omega(G)$  has bounded difference property with  $L = 1$ .

- ▶ For two graphs  $G$  and  $G'$  that differ in at most 1 edge,

$$|\omega(G) - \omega(G')| \leq 1.$$

- ▶ Thus  $E(G) \rightarrow \omega(G)$  has bounded difference property with  $L = 1$ .
- ▶ Let  $G$  be an Erdős-Rényi random graph: Edges are independently drawn with probability  $p$ .

- ▶ For two graphs  $G$  and  $G'$  that differ in at most 1 edge,

$$|\omega(G) - \omega(G')| \leq 1.$$

- ▶ Thus  $E(G) \rightarrow \omega(G)$  has bounded difference property with  $L = 1$ .
- ▶ Let  $G$  be an Erdős-Rényi random graph: Edges are independently drawn with probability  $p$ . Then, with  $m = \binom{n}{2}$

$$\mathbb{P}(|\omega(G) - \mathbb{E}(\omega(G))| \geq \delta) \leq 2 \exp(-2\delta^2 / \sum_{i=1}^m 1)$$

- ▶ For two graphs  $G$  and  $G'$  that differ in at most 1 edge,

$$|\omega(G) - \omega(G')| \leq 1.$$

- ▶ Thus  $E(G) \rightarrow \omega(G)$  has bounded difference property with  $L = 1$ .
- ▶ Let  $G$  be an Erdős-Rényi random graph: Edges are independently drawn with probability  $p$ . Then, with  $m = \binom{n}{2}$

$$\begin{aligned} \mathbb{P}(|\omega(G) - \mathbb{E}(\omega(G))| \geq \delta) &\leq 2 \exp(-2\delta^2 / \sum_{i=1}^m 1) \\ &= 2 \exp(-2\delta^2 / m). \end{aligned}$$



- ▶ For two graphs  $G$  and  $G'$  that differ in at most 1 edge,

$$|\omega(G) - \omega(G')| \leq 1.$$

- ▶ Thus  $E(G) \rightarrow \omega(G)$  has bounded difference property with  $L = 1$ .
- ▶ Let  $G$  be an Erdős-Rényi random graph: Edges are independently drawn with probability  $p$ . Then, with  $m = \binom{n}{2}$

$$\begin{aligned} \mathbb{P}(|\omega(G) - \mathbb{E}(\omega(G))| \geq \delta) &\leq 2 \exp(-2\delta^2 / \sum_{i=1}^m 1) \\ &= 2 \exp(-2\delta^2 / m). \end{aligned}$$

or setting  $\bar{\omega}(G) = \omega(G)/m$ ,

$$\mathbb{P}(|\bar{\omega}(G) - \mathbb{E}(\bar{\omega}(G))| \geq \delta) = \mathbb{P}(|\omega(G) - \mathbb{E}(\omega(G))| \geq m\delta)$$

- ▶ For two graphs  $G$  and  $G'$  that differ in at most 1 edge,

$$|\omega(G) - \omega(G')| \leq 1.$$

- ▶ Thus  $E(G) \rightarrow \omega(G)$  has bounded difference property with  $L = 1$ .
- ▶ Let  $G$  be an Erdős-Rényi random graph: Edges are independently drawn with probability  $p$ . Then, with  $m = \binom{n}{2}$

$$\begin{aligned} \mathbb{P}(|\omega(G) - \mathbb{E}(\omega(G))| \geq \delta) &\leq 2 \exp(-2\delta^2 / \sum_{i=1}^m 1) \\ &= 2 \exp(-2\delta^2 / m). \end{aligned}$$

or setting  $\bar{\omega}(G) = \omega(G)/m$ ,

$$\begin{aligned} \mathbb{P}(|\bar{\omega}(G) - \mathbb{E}(\bar{\omega}(G))| \geq \delta) &= \mathbb{P}(|\omega(G) - \mathbb{E}(\omega(G))| \geq m\delta) \\ &\leq 2 \exp(-2(m\delta)^2 / m) \end{aligned}$$

- ▶ For two graphs  $G$  and  $G'$  that differ in at most 1 edge,

$$|\omega(G) - \omega(G')| \leq 1.$$

- ▶ Thus  $E(G) \rightarrow \omega(G)$  has bounded difference property with  $L = 1$ .
- ▶ Let  $G$  be an Erdős-Rényi random graph: Edges are independently drawn with probability  $p$ . Then, with  $m = \binom{n}{2}$

$$\begin{aligned} \mathbb{P}(|\omega(G) - \mathbb{E}(\omega(G))| \geq \delta) &\leq 2 \exp(-2\delta^2 / \sum_{i=1}^m 1) \\ &= 2 \exp(-2\delta^2 / m). \end{aligned}$$

or setting  $\bar{\omega}(G) = \omega(G)/m$ ,

$$\begin{aligned} \mathbb{P}(|\bar{\omega}(G) - \mathbb{E}(\bar{\omega}(G))| \geq \delta) &= \mathbb{P}(|\omega(G) - \mathbb{E}(\omega(G))| \geq m\delta) \\ &\leq 2 \exp(-2(m\delta)^2 / m) \\ &\leq 2 \exp(-2m\delta^2) \end{aligned}$$

# Lipschitz functions of standard Gaussian vector

- ▶ A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L$ -Lipschitz w.r.t.  $\|\cdot\|_2$  if

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

# Lipschitz functions of standard Gaussian vector

- ▶ A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L$ -Lipschitz w.r.t.  $\|\cdot\|_2$  if

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

- ▶ For instance, if  $f$  is differentiable with bounded derivative.

# Lipschitz functions of standard Gaussian vector

- ▶ A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L$ -Lipschitz w.r.t.  $\|\cdot\|_2$  if

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

- ▶ For instance, if  $f$  is differentiable with bounded derivative.
- ▶ The function  $f : \mathbb{R} \rightarrow \mathbb{R}$ ,  $f(x) = e^{-x^2}$  is Lipschitz.

# Lipschitz functions of standard Gaussian vector

- ▶ A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L$ -Lipschitz w.r.t.  $\|\cdot\|_2$  if

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

- ▶ For instance, if  $f$  is differentiable with bounded derivative.
- ▶ The function  $f : \mathbb{R} \rightarrow \mathbb{R}$ ,  $f(x) = e^{-x^2}$  is Lipschitz.
- ▶ The function  $f : [0, 1] \rightarrow \mathbb{R}$ ,  $f(x) = \sqrt{x}$  is not Lipschitz.

- **Theorem 9 (Gaussian concentration).** Let  $X \sim N(0, I_n)$  be a standard Gaussian vector and assume that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L$ -Lipschitz w.r.t. the Euclidean norm. Then,

$$\mathbb{P}(f(X) - \mathbb{E}(f(X)) \geq t) \leq \exp\left(-\frac{t^2}{2L^2}\right), \quad t \geq 0.$$



- **Theorem 9 (Gaussian concentration).** Let  $X \sim N(0, I_n)$  be a standard Gaussian vector and assume that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L$ -Lipschitz w.r.t. the Euclidean norm. Then,

$$\mathbb{P}(f(X) - \mathbb{E}(f(X)) \geq t) \leq \exp\left(-\frac{t^2}{2L^2}\right), \quad t \geq 0.$$

- Notice that it follows that

$$\mathbb{P}(|f(X) - \mathbb{E}(f(X))| \geq t) \leq 2 \exp\left(-\frac{t^2}{2L^2}\right), \quad t \geq 0.$$

- **Theorem 9 (Gaussian concentration).** Let  $X \sim N(0, I_n)$  be a standard Gaussian vector and assume that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L$ -Lipschitz w.r.t. the Euclidean norm. Then,

$$\mathbb{P}(f(X) - \mathbb{E}(f(X)) \geq t) \leq \exp\left(-\frac{t^2}{2L^2}\right), \quad t \geq 0.$$

- Notice that it follows that

$$\mathbb{P}(|f(X) - \mathbb{E}(f(X))| \geq t) \leq 2 \exp\left(-\frac{t^2}{2L^2}\right), \quad t \geq 0.$$

- In other words,  $f(X)$  is sub-Gaussian with parameter  $L$ .

- **Theorem 9 (Gaussian concentration).** Let  $X \sim N(0, I_n)$  be a standard Gaussian vector and assume that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L$ -Lipschitz w.r.t. the Euclidean norm. Then,

$$\mathbb{P}(f(X) - \mathbb{E}(f(X)) \geq t) \leq \exp\left(-\frac{t^2}{2L^2}\right), \quad t \geq 0.$$

- Notice that it follows that

$$\mathbb{P}(|f(X) - \mathbb{E}(f(X))| \geq t) \leq 2 \exp\left(-\frac{t^2}{2L^2}\right), \quad t \geq 0.$$

- In other words,  $f(X)$  is sub-Gaussian with parameter  $L$ .
- Deep result, no easy proof!

- **Theorem 9 (Gaussian concentration).** Let  $X \sim N(0, I_n)$  be a standard Gaussian vector and assume that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $L$ -Lipschitz w.r.t. the Euclidean norm. Then,

$$\mathbb{P}(f(X) - \mathbb{E}(f(X)) \geq t) \leq \exp\left(-\frac{t^2}{2L^2}\right), \quad t \geq 0.$$

- Notice that it follows that

$$\mathbb{P}(|f(X) - \mathbb{E}(f(X))| \geq t) \leq 2 \exp\left(-\frac{t^2}{2L^2}\right), \quad t \geq 0.$$

- In other words,  $f(X)$  is sub-Gaussian with parameter  $L$ .
- Deep result, no easy proof!
- Has far-reaching consequences.

## Example: Singular values

- ▶ Consider a matrix  $X \in \mathbb{R}^{n \times d}$  where  $n > d$ .

## Example: Singular values

- ▶ Consider a matrix  $X \in \mathbb{R}^{n \times d}$  where  $n > d$ .
- ▶ Let  $\sigma_1(X) \geq \sigma_2(X) \geq \dots \geq \sigma_k(X)$  be (ordered) singular values of  $X$ .

## Example: Singular values

- ▶ Consider a matrix  $X \in \mathbb{R}^{n \times d}$  where  $n > d$ .
- ▶ Let  $\sigma_1(X) \geq \sigma_2(X) \geq \dots \geq \sigma_k(X)$  be (ordered) singular values of  $X$ .
- ▶ By Weyl's theorem, for any  $X, Y \in \mathbb{R}^{n \times d}$ :

$$|\sigma_k(X) - \sigma_k(Y)| \leq \|X - Y\|_{op} \leq \|X - Y\|_F.$$

Note that this is a generalization of order-statistics inequality.)

## Example: Singular values

- ▶ Consider a matrix  $X \in \mathbb{R}^{n \times d}$  where  $n > d$ .
- ▶ Let  $\sigma_1(X) \geq \sigma_2(X) \geq \dots \geq \sigma_k(X)$  be (ordered) singular values of  $X$ .
- ▶ By Weyl's theorem, for any  $X, Y \in \mathbb{R}^{n \times d}$ :

$$|\sigma_k(X) - \sigma_k(Y)| \leq \|X - Y\|_{op} \leq \|X - Y\|_F.$$

Note that this is a generalization of order-statistics inequality.)

- ▶ Thus,  $X \rightarrow \sigma_1(X)$  is 1-Lipschitz.



## Example: Singular values

- ▶ Consider a matrix  $X \in \mathbb{R}^{n \times d}$  where  $n > d$ .
- ▶ Let  $\sigma_1(X) \geq \sigma_2(X) \geq \dots \geq \sigma_k(X)$  be (ordered) singular values of  $X$ .
- ▶ By Weyl's theorem, for any  $X, Y \in \mathbb{R}^{n \times d}$ :

$$|\sigma_k(X) - \sigma_k(Y)| \leq \|X - Y\|_{op} \leq \|X - Y\|_F.$$

Note that this is a generalization of order-statistics inequality.)

- ▶ Thus,  $X \rightarrow \sigma_1(X)$  is 1-Lipschitz.
- ▶ **Proposition 6.** Let  $X \in \mathbb{R}^{n \times d}$  be a random matrix with iid  $N(0, 1)$  entries. Then,

$$\mathbb{P}(|\sigma_k(X) - \mathbb{E}(\sigma_k(X))| \geq \delta) \leq 2e^{-\delta^2/2}, \quad \delta \geq 0.$$

## Example: Singular values

- ▶ Consider a matrix  $X \in \mathbb{R}^{n \times d}$  where  $n > d$ .
- ▶ Let  $\sigma_1(X) \geq \sigma_2(X) \geq \dots \geq \sigma_k(X)$  be (ordered) singular values of  $X$ .
- ▶ By Weyl's theorem, for any  $X, Y \in \mathbb{R}^{n \times d}$ :

$$|\sigma_k(X) - \sigma_k(Y)| \leq \|X - Y\|_{op} \leq \|X - Y\|_F.$$

Note that this is a generalization of order-statistics inequality.)

- ▶ Thus,  $X \rightarrow \sigma_1(X)$  is 1-Lipschitz.
- ▶ **Proposition 6.** Let  $X \in \mathbb{R}^{n \times d}$  be a random matrix with iid  $N(0, 1)$  entries. Then,

$$\mathbb{P}(|\sigma_k(X) - \mathbb{E}(\sigma_k(X))| \geq \delta) \leq 2e^{-\delta^2/2}, \quad \delta \geq 0.$$

- ▶ It remains to characterize  $\mathbb{E}(\sigma_k(X))$ .

## Sparsity models

- ▶ When  $d > n$  there is no hope of estimating  $\theta^*$ ,

## Sparsity models

- ▶ When  $d > n$  there is no hope of estimating  $\theta^*$ ,
- ▶ unless we impose a low dimensional assumption on  $\theta^*$ .

## Sparsity models

- ▶ When  $d > n$  there is no hope of estimating  $\theta^*$ ,
- ▶ unless we impose a low dimensional assumption on  $\theta^*$ .
- ▶ Support of  $\theta^*$  (recall  $[d] = \{1, \dots, d\}$ )

$$\text{supp}(\theta^*) := S(\theta^*) = \{j \in [d] : \theta_j^* \neq 0\}.$$

# Sparsity models

- ▶ When  $d > n$  there is no hope of estimating  $\theta^*$ ,
- ▶ unless we impose a low dimensional assumption on  $\theta^*$ .
- ▶ Support of  $\theta^*$  (recall  $[d] = \{1, \dots, d\}$ )

$$\text{supp}(\theta^*) := S(\theta^*) = \{j \in [d] : \theta_j^* \neq 0\}.$$

- ▶ Hard sparsity assumption:  $s = |S(\theta^*)| \ll d$ .

# Sparsity models

- ▶ When  $d > n$  there is no hope of estimating  $\theta^*$ ,
- ▶ unless we impose a low dimensional assumption on  $\theta^*$ .
- ▶ Support of  $\theta^*$  (recall  $[d] = \{1, \dots, d\}$ )

$$\text{supp}(\theta^*) := S(\theta^*) = \{j \in [d] : \theta_j^* \neq 0\}.$$

- ▶ Hard sparsity assumption:  $s = |S(\theta^*)| \ll d$ .
- ▶ Weak sparsity via  $\ell_q$  balls for  $q \in [0, 1]$ :

$$\theta^* \in \mathbb{B}_q(R_q) := \left\{ \theta \in \mathbb{R}^d : \sum_{j=1}^d |\theta_j|^q \leq R_q \right\}.$$

# Sparsity models

- ▶ When  $d > n$  there is no hope of estimating  $\theta^*$ ,
- ▶ unless we impose a low dimensional assumption on  $\theta^*$ .
- ▶ Support of  $\theta^*$  (recall  $[d] = \{1, \dots, d\}$ )

$$\text{supp}(\theta^*) := S(\theta^*) = \{j \in [d] : \theta_j^* \neq 0\}.$$

- ▶ Hard sparsity assumption:  $s = |S(\theta^*)| \ll d$ .
- ▶ Weak sparsity via  $\ell_q$  balls for  $q \in [0, 1]$ :

$$\theta^* \in \mathbb{B}_q(R_q) := \left\{ \theta \in \mathbb{R}^d : \sum_{j=1}^d |\theta_j|^q \leq R_q \right\}.$$

- ▶  $q = 1$  gives the  $\ell_1$  ball.



# Sparsity models

- ▶ When  $d > n$  there is no hope of estimating  $\theta^*$ ,
- ▶ unless we impose a low dimensional assumption on  $\theta^*$ .
- ▶ Support of  $\theta^*$  (recall  $[d] = \{1, \dots, d\}$ )

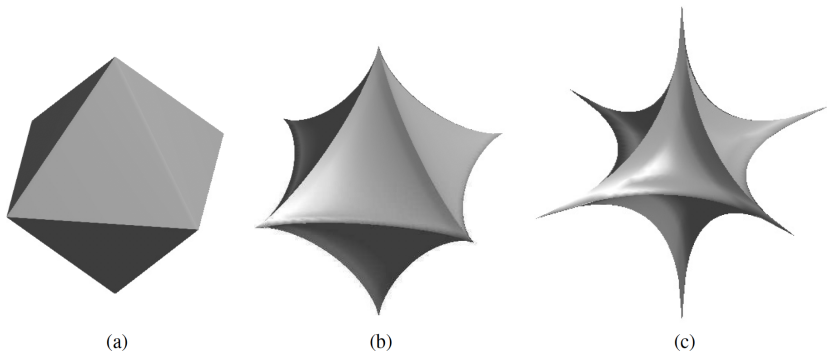
$$\text{supp}(\theta^*) := S(\theta^*) = \{j \in [d] : \theta_j^* \neq 0\}.$$

- ▶ Hard sparsity assumption:  $s = |S(\theta^*)| \ll d$ .
- ▶ Weak sparsity via  $\ell_q$  balls for  $q \in [0, 1]$ :

$$\theta^* \in \mathbb{B}_q(R_q) := \left\{ \theta \in \mathbb{R}^d : \sum_{j=1}^d |\theta_j|^q \leq R_q \right\}.$$

- ▶  $q = 1$  gives the  $\ell_1$  ball.
- ▶  $q = 0$  gives the  $\ell_0$  ball, same as hard sparsity:

$$\|\theta^*\|_0 := |S(\theta^*)|.$$



**Figure 7.1** Illustrations of the  $\ell_q$ -“balls” for different choices of the parameter  $q \in (0, 1]$ . (a) For  $q = 1$ , the set  $\mathbb{B}_1(R_q)$  corresponds to the usual  $\ell_1$ -ball shown here. (b) For  $q = 0.75$ , the ball is a non-convex set obtained by collapsing the faces of the  $\ell_1$ -ball towards the origin. (c) For  $q = 0.5$ , the set becomes more “spiky”, and it collapses into the hard sparsity constraint as  $q \rightarrow 0^+$ . As shown in Exercise 7.2(a), for all  $q \in (0, 1]$ , the set  $\mathbb{B}_q(1)$  is star-shaped around the origin.

# Compressed sensing or the equation $y = X\theta^*$

- ▶ When can we solve the equation  $y = X\theta^*$ ?

## The classical answer

The classical theory of linear algebra, which we learn as undergraduates, is as follows:

- ▶ If there are at least as many equations as unknowns ( $n \geq d$ ), and  $X$  has full rank, then the problem is determined or overdetermined, and one can easily solve  $y = X\theta$  uniquely (e.g. by gaussian elimination).

# The classical answer

The classical theory of linear algebra, which we learn as undergraduates, is as follows:

- ▶ If there are at least as many equations as unknowns ( $n \geq d$ ), and  $X$  has full rank, then the problem is determined or overdetermined, and one can easily solve  $y = X\theta$  uniquely (e.g. by gaussian elimination).
- ▶ If there are fewer equations than unknowns ( $n < d$ ), then the problem is underdetermined even when  $X$  has full rank. Knowledge of  $y = X\theta$  restricts  $\theta$  to an (affine) subspace of  $\mathbb{R}^d$ , but does not determine  $\theta$  completely.

# Sparse recovery

- ▶ It is thus of interest to obtain a good estimator for underdetermined problems such as  $X\theta^* = y$  in the case in which  $\theta^*$  is expected to be “spiky” - that is, concentrated in only a few of its coordinates.

# Sparse recovery

- ▶ It is thus of interest to obtain a good estimator for underdetermined problems such as  $X\theta^* = y$  in the case in which  $\theta^*$  is expected to be “spiky” - that is, concentrated in only a few of its coordinates.
- ▶ A model case occurs when  $\theta^*$  is known to be  $s$ -sparse for some  $1 \leq s \leq d$ , which means that at most  $s$  of the coefficients of  $\theta^*$  can be non-zero.

# Sparse recovery

- ▶ Sparsity is a simple but effective model for many real-life signals. For instance, an image may be many megapixels in size, but when viewed in the right basis (e.g. a wavelet basis), many of the coefficients may be negligible, and so the image may be compressible into a file of much smaller size without seriously affecting the image quality. In other words, many images are effectively sparse in the wavelet basis.



# Sparsity helps!

- ▶ Intuitively, if a signal  $\theta^* \in \mathbb{R}^d$  is  $s$ -sparse, then it should only have  $s$  degrees of freedom rather than  $d$ . In principle, one should now only need  $s$  measurements or so to reconstruct  $\theta^*$ , rather than  $d$ .

Compressed sensing is advantageous whenever:

- ▶ signals are sparse in a known basis;

Compressed sensing is advantageous whenever:

- ▶ signals are sparse in a known basis;
- ▶ measurements (or computation at the sensor end) are expensive; but computations at the receiver end are cheap.

Compressed sensing is advantageous whenever:

- ▶ signals are sparse in a known basis;
- ▶ measurements (or computation at the sensor end) are expensive; but computations at the receiver end are cheap.

Compressed sensing is advantageous whenever:

- ▶ signals are sparse in a known basis;
- ▶ measurements (or computation at the sensor end) are expensive; but computations at the receiver end are cheap.

Such situations can arise in:

Compressed sensing is advantageous whenever:

- ▶ signals are sparse in a known basis;
- ▶ measurements (or computation at the sensor end) are expensive; but computations at the receiver end are cheap.

Such situations can arise in:

- ▶ Imaging.

Compressed sensing is advantageous whenever:

- ▶ signals are sparse in a known basis;
- ▶ measurements (or computation at the sensor end) are expensive; but computations at the receiver end are cheap.

Such situations can arise in:

- ▶ Imaging.
- ▶ Sensor networks.

Compressed sensing is advantageous whenever:

- ▶ signals are sparse in a known basis;
- ▶ measurements (or computation at the sensor end) are expensive; but computations at the receiver end are cheap.

Such situations can arise in:

- ▶ Imaging.
- ▶ Sensor networks.
- ▶ MRI.



Compressed sensing is advantageous whenever:

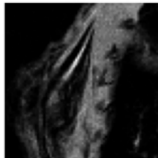
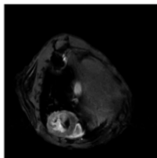
- ▶ signals are sparse in a known basis;
- ▶ measurements (or computation at the sensor end) are expensive; but computations at the receiver end are cheap.

Such situations can arise in:

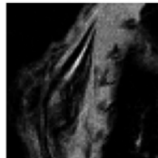
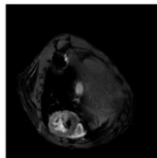
- ▶ Imaging.
- ▶ Sensor networks.
- ▶ MRI.
- ▶ Astronomy.

# Sparsity helps

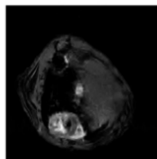
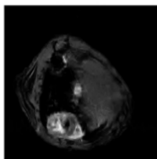
Fully sampled



Fully sampled



Sparse reconstruction from under-sampled observations    Sparse reconstruction from under-sampled observations



# Basis pursuit

- ▶ Consider the noiseless case  $y = X\theta^*$ .

# Basis pursuit

- ▶ Consider the noiseless case  $y = X\theta^*$ .
- ▶ We assume that  $\|\theta^*\|_0$  is small.

# Basis pursuit

- ▶ Consider the noiseless case  $y = X\theta^*$ .
- ▶ We assume that  $\|\theta^*\|_0$  is small.
- ▶ Ideal program to solve:

$$\min_{\theta} \|\theta\|_0 \quad \text{subject to} \quad y = X\theta$$

# Basis pursuit

- ▶ Consider the noiseless case  $y = X\theta^*$ .
- ▶ We assume that  $\|\theta^*\|_0$  is small.
- ▶ Ideal program to solve:

$$\min_{\theta} \|\theta\|_0 \quad \text{subject to } y = X\theta$$

- ▶  $\|\cdot\|_0$  is highly nonconvex, relax to  $\|\cdot\|_1$ :

$$\min_{\theta} \|\theta\|_1 \quad \text{subject to } y = X\theta$$

# Basis pursuit

- ▶ Consider the noiseless case  $y = X\theta^*$ .
- ▶ We assume that  $\|\theta^*\|_0$  is small.
- ▶ Ideal program to solve:

$$\min_{\theta} \|\theta\|_0 \quad \text{subject to } y = X\theta$$

- ▶  $\|\cdot\|_0$  is highly nonconvex, relax to  $\|\cdot\|_1$ :

$$\min_{\theta} \|\theta\|_1 \quad \text{subject to } y = X\theta$$

- ▶ This is called basis pursuit (regression).

# Basis pursuit

- ▶ Consider the noiseless case  $y = X\theta^*$ .
- ▶ We assume that  $\|\theta^*\|_0$  is small.
- ▶ Ideal program to solve:

$$\min_{\theta} \|\theta\|_0 \quad \text{subject to } y = X\theta$$

- ▶  $\|\cdot\|_0$  is highly nonconvex, relax to  $\|\cdot\|_1$ :

$$\min_{\theta} \|\theta\|_1 \quad \text{subject to } y = X\theta$$

- ▶ This is called basis pursuit (regression).
- ▶ The resulting problem is convex.



# Basis pursuit

- In fact, can be written as a linear program. Notice that the problem is equivalent to

$$\min_s \sum_{j=1}^d s_j \quad \text{subject to} \quad y = X\theta, \quad |\theta_j| \leq s_j.$$

# Basis pursuit

- In fact, can be written as a linear program. Notice that the problem is equivalent to

$$\min_s \sum_{j=1}^d s_j \quad \text{subject to} \quad y = X\theta, \quad |\theta_j| \leq s_j.$$

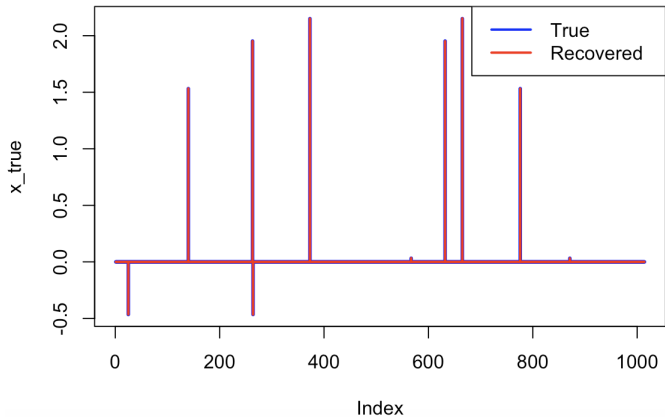
- Global solutions can be obtained very efficiently.

# Basis pursuit example

```
1 library(CVXR)
2
3
4 set.seed(123)
5 n <- 1014           # Signal dimension
6 m <- 64             # Number of measurements
7 A <- matrix(rnorm(m * n), m, n) # Measurement matrix
8
9
10
11 x_true <- rep(0, n)
12 x_true[sample(1:n, 10)] <- rnorm(5) # Sparse true signal
13 plot(x_true)
14 b <- A %*% x_true
15
16
17 x <- Variable(n)
18 objective <- Minimize(norm1(x))
19 constraints <- list(A %*% x == b)
20 problem <- Problem(objective, constraints)
21
22 result <- solve(problem)
23 x_est <- result$getValue(x)
24
25
26 # Plot results
27 plot(x_true, type = "h", lwd = 3, col = "blue", ylim = range(c(x_true, x_est)),
28      main = "True vs Recovered Signal")
29 lines(x_est, type = "h", col = "red", lwd = 2)
30 legend("topright", legend = c("True", "Recovered"), col = c("blue", "red"), lwd = 2)
31
32 mean((x_est - x_true)^2)
33 - #####
```

# Basis pursuit example

**True vs Recovered Signal**



Terence Tao and Emmanuel Candes have been celebrated for understanding of compressed sensing



# Restricted null space property (RNS)

► Define

$$\mathbb{C}(\mathbf{S}) := \{\Delta \in \mathbb{R}^d : \|\Delta_{\mathbf{S}^c}\|_1 \leq \|\Delta_{\mathbf{S}}\|_1\}.$$

# Restricted null space property (RNS)

► Define

$$\mathbb{C}(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}.$$

## Theorem

*The following two are equivalent:*

# Restricted null space property (RNS)

- ▶ Define

$$\mathbb{C}(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}.$$

## Theorem

*The following two are equivalent:*

- ▶ *For any  $\theta^* \in \mathbb{R}^d$  with support  $\subset S$ , the basis pursuit program applied to the data  $(X, y = X\theta^*)$  has unique solution  $\hat{\theta} = \theta^*$ .*



# Restricted null space property (RNS)

- ▶ Define

$$\mathbb{C}(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}.$$

## Theorem

*The following two are equivalent:*

- ▶ *For any  $\theta^* \in \mathbb{R}^d$  with support  $\subset S$ , the basis pursuit program applied to the data  $(X, y = X\theta^*)$  has unique solution  $\hat{\theta} = \theta^*$ .*
- ▶ *The restricted null space (RNS) property holds, i.e.,*

$$\mathbb{C}(S) \cap \ker(X) = \{0\}.$$

*Recall that  $\ker(X) = \{\theta \in \mathbb{R}^d : X\theta = 0\}$ .*

## Proof

- Consider the tangent cone to the  $\ell_1$  ball (of radius  $\|\theta^*\|_1$ ) at  $\theta^*$ :

$$\mathbb{T}(\theta^*) = \{\Delta \in \mathbb{R}^d : \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \text{ for some } t > 0\}.$$

## Proof

- Consider the tangent cone to the  $\ell_1$  ball (of radius  $\|\theta^*\|_1$ ) at  $\theta^*$ :

$$\mathbb{T}(\theta^*) = \{\Delta \in \mathbb{R}^d : \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \text{ for some } t > 0\}.$$

i.e., the set of descent directions for the norm  $\ell_1$  at the point  $\theta^*$ .

## Proof

- ▶ Consider the tangent cone to the  $\ell_1$  ball (of radius  $\|\theta^*\|_1$ ) at  $\theta^*$ :

$$\mathbb{T}(\theta^*) = \{\Delta \in \mathbb{R}^d : \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \text{ for some } t > 0\}.$$

i.e., the set of descent directions for the norm  $\ell_1$  at the point  $\theta^*$ .

- ▶ Feasible set is  $\theta^* + \ker(X)$ , i.e.,  $\ker(X)$  is the set of feasible directions  $\Delta = \theta - \theta^*$ .

## Proof

- ▶ Consider the tangent cone to the  $\ell_1$  ball (of radius  $\|\theta^*\|_1$ ) at  $\theta^*$ :

$$\mathbb{T}(\theta^*) = \{\Delta \in \mathbb{R}^d : \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \text{ for some } t > 0\}.$$

i.e., the set of descent directions for the norm  $\ell_1$  at the point  $\theta^*$ .

- ▶ Feasible set is  $\theta^* + \ker(X)$ , i.e.,  $\ker(X)$  is the set of feasible directions  $\Delta = \theta - \theta^*$ .
- ▶ Hence, there is a minimizer other than  $\theta^*$  if and only if

$$\mathbb{T}(\theta^*) \cap \ker(X) \neq \{0\}.$$

## Proof

- ▶ Consider the tangent cone to the  $\ell_1$  ball (of radius  $\|\theta^*\|_1$ ) at  $\theta^*$ :

$$\mathbb{T}(\theta^*) = \{\Delta \in \mathbb{R}^d : \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \text{ for some } t > 0\}.$$

i.e., the set of descent directions for the norm  $\ell_1$  at the point  $\theta^*$ .

- ▶ Feasible set is  $\theta^* + \ker(X)$ , i.e.,  $\ker(X)$  is the set of feasible directions  $\Delta = \theta - \theta^*$ .
- ▶ Hence, there is a minimizer other than  $\theta^*$  if and only if

$$\mathbb{T}(\theta^*) \cap \ker(X) \neq \{0\}.$$

And so the minimizer is  $\theta^*$  if and only if  $\mathbb{T}(\theta^*) \cap \ker(X) = \{0\}$ .

## Proof

- Consider the tangent cone to the  $\ell_1$  ball (of radius  $\|\theta^*\|_1$ ) at  $\theta^*$ :

$$\mathbb{T}(\theta^*) = \{\Delta \in \mathbb{R}^d : \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \text{ for some } t > 0\}.$$

i.e., the set of descent directions for the norm  $\ell_1$  at the point  $\theta^*$ .

- Feasible set is  $\theta^* + \ker(X)$ , i.e.,  $\ker(X)$  is the set of feasible directions  $\Delta = \theta - \theta^*$ .
- Hence, there is a minimizer other than  $\theta^*$  if and only if

$$\mathbb{T}(\theta^*) \cap \ker(X) \neq \{0\}.$$

And so the minimizer is  $\theta^*$  if and only if  $\mathbb{T}(\theta^*) \cap \ker(X) = \{0\}$ .

- It is enough to show that

$$\mathbb{C}(\mathcal{S}) = \bigcup_{\theta \in \mathbb{R}^d : \text{supp}(\theta) \subset \mathcal{S}} \mathbb{T}(\theta)$$

## Proof

- Consider the tangent cone to the  $\ell_1$  ball (of radius  $\|\theta^*\|_1$ ) at  $\theta^*$ :

$$\mathbb{T}(\theta^*) = \{\Delta \in \mathbb{R}^d : \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \text{ for some } t > 0\}.$$

i.e., the set of descent directions for the norm  $\ell_1$  at the point  $\theta^*$ .

- Feasible set is  $\theta^* + \ker(X)$ , i.e.,  $\ker(X)$  is the set of feasible directions  $\Delta = \theta - \theta^*$ .
- Hence, there is a minimizer other than  $\theta^*$  if and only if

$$\mathbb{T}(\theta^*) \cap \ker(X) \neq \{0\}.$$

And so the minimizer is  $\theta^*$  if and only if  $\mathbb{T}(\theta^*) \cap \ker(X) = \{0\}$ .

- It is enough to show that

$$\mathbb{C}(S) = \bigcup_{\theta \in \mathbb{R}^d : \text{supp}(\theta) \subset S} \mathbb{T}(\theta)$$

since then

$$\mathbb{C}(S) \cap \ker(X) = \bigcup_{\theta \in \mathbb{R}^d : \text{supp}(\theta) \subset S} (\mathbb{T}(\theta) \cap \ker(X)).$$



- It is enough to show that

$$\mathbb{C}(S) = \bigcup_{\theta \in \mathbb{R}^d : \text{supp}(\theta) \subset S} \mathbb{T}(\theta)$$

- It is enough to show that

$$\mathbb{C}(S) = \bigcup_{\theta \in \mathbb{R}^d : \text{supp}(\theta) \subset S} \mathbb{T}(\theta)$$

- Let  $\mathbb{T}_1(\theta^*)$  be the subset of  $\mathbb{T}(\theta^*)$  with  $t = 1$ . We can work with  $\mathbb{T}_1(\theta^*)$  instead of  $\mathbb{T}(\theta^*)$ .

- It is enough to show that

$$\mathbb{C}(S) = \bigcup_{\theta \in \mathbb{R}^d : \text{supp}(\theta) \subset S} \mathbb{T}(\theta)$$

- Let  $\mathbb{T}_1(\theta^*)$  be the subset of  $\mathbb{T}(\theta^*)$  with  $t = 1$ . We can work with  $\mathbb{T}_1(\theta^*)$  instead of  $\mathbb{T}(\theta^*)$ .
- Notice that

$$\|\Delta + \theta^*\|_1 = \|\Delta_{S^c} + \theta_{S^c}^*\|_1 + \|\Delta_S + \theta_S^*\|_1$$

- It is enough to show that

$$\mathbb{C}(S) = \bigcup_{\theta \in \mathbb{R}^d : \text{supp}(\theta) \subset S} \mathbb{T}(\theta)$$

- Let  $\mathbb{T}_1(\theta^*)$  be the subset of  $\mathbb{T}(\theta^*)$  with  $t = 1$ . We can work with  $\mathbb{T}_1(\theta^*)$  instead of  $\mathbb{T}(\theta^*)$ .
- Notice that

$$\|\Delta + \theta^*\|_1 = \|\Delta_{S^c} + \theta_{S^c}^*\|_1 + \|\Delta_S + \theta_S^*\|_1 = \|\Delta_{S^c}\|_1 + \|\Delta_S + \theta_S^*\|_1.$$

- It is enough to show that

$$\mathbb{C}(S) = \bigcup_{\theta \in \mathbb{R}^d : \text{supp}(\theta) \subset S} \mathbb{T}(\theta)$$

- Let  $\mathbb{T}_1(\theta^*)$  be the subset of  $\mathbb{T}(\theta^*)$  with  $t = 1$ . We can work with  $\mathbb{T}_1(\theta^*)$  instead of  $\mathbb{T}(\theta^*)$ .
- Notice that

$$\|\Delta + \theta^*\|_1 = \|\Delta_{S^c} + \theta_{S^c}^*\|_1 + \|\Delta_S + \theta_S^*\|_1 = \|\Delta_{S^c}\|_1 + \|\Delta_S + \theta_S^*\|_1.$$

Hence,  $\Delta \in \mathbb{T}_1(\theta^*)$  if and only if

$$\|\Delta_{S^c}\|_1 + \|\Delta_S + \theta_S^*\|_1 = \|\Delta + \theta^*\|_1$$

- It is enough to show that

$$\mathbb{C}(S) = \bigcup_{\theta \in \mathbb{R}^d : \text{supp}(\theta) \subset S} \mathbb{T}(\theta)$$

- Let  $\mathbb{T}_1(\theta^*)$  be the subset of  $\mathbb{T}(\theta^*)$  with  $t = 1$ . We can work with  $\mathbb{T}_1(\theta^*)$  instead of  $\mathbb{T}(\theta^*)$ .
- Notice that

$$\|\Delta + \theta^*\|_1 = \|\Delta_{S^c} + \theta_{S^c}^*\|_1 + \|\Delta_S + \theta_S^*\|_1 = \|\Delta_{S^c}\|_1 + \|\Delta_S + \theta_S^*\|_1.$$

Hence,  $\Delta \in \mathbb{T}_1(\theta^*)$  if and only if

$$\|\Delta_{S^c}\|_1 + \|\Delta_S + \theta_S^*\|_1 = \|\Delta + \theta^*\|_1 \leq \|\theta^*\|_1 = \|\theta_S^*\|_1$$

- It is enough to show that

$$\mathbb{C}(S) = \bigcup_{\theta \in \mathbb{R}^d : \text{supp}(\theta) \subset S} \mathbb{T}(\theta)$$

- Let  $\mathbb{T}_1(\theta^*)$  be the subset of  $\mathbb{T}(\theta^*)$  with  $t = 1$ . We can work with  $\mathbb{T}_1(\theta^*)$  instead of  $\mathbb{T}(\theta^*)$ .
- Notice that

$$\|\Delta + \theta^*\|_1 = \|\Delta_{S^c} + \theta_{S^c}^*\|_1 + \|\Delta_S + \theta_S^*\|_1 = \|\Delta_{S^c}\|_1 + \|\Delta_S + \theta_S^*\|_1.$$

Hence,  $\Delta \in \mathbb{T}_1(\theta^*)$  if and only if

$$\|\Delta_{S^c}\|_1 + \|\Delta_S + \theta_S^*\|_1 = \|\Delta + \theta^*\|_1 \leq \|\theta^*\|_1 = \|\theta_S^*\|_1$$

if and only if

$$\|\Delta_{S^c}\|_1 \leq \|\theta_S^*\|_1 - \|\Delta_S + \theta_S^*\|_1.$$

- We have shown that  $\Delta \in \mathbb{T}_1(\theta^*)$  if and only if  $\|\Delta_{S^c}\|_1 \leq \|\theta_S^*\|_1 - \|\Delta_S + \theta_S^*\|_1$ .



- ▶ We have shown that  $\Delta \in \mathbb{T}_1(\theta^*)$  if and only if  $\|\Delta_{S^c}\|_1 \leq \|\theta_S^*\|_1 - \|\Delta_S + \theta_S^*\|_1$ .
- ▶ We have that  $\Delta \in \mathbb{T}_1(\theta^*)$  for some  $\theta^*$  with  $\text{supp}(\theta^*) \subset S$  iff

$$\|\Delta_{S^c}\|_1 \leq \sup_{\theta^* \in \mathbb{R}^d} \{\|\theta_S^*\|_1 - \|\Delta_S + \theta_S^*\|_1\}$$

- ▶ We have shown that  $\Delta \in \mathbb{T}_1(\theta^*)$  if and only if  $\|\Delta_{S^c}\|_1 \leq \|\theta_S^*\|_1 - \|\Delta_S + \theta_S^*\|_1$ .
- ▶ We have that  $\Delta \in \mathbb{T}_1(\theta^*)$  for some  $\theta^*$  with  $\text{supp}(\theta^*) \subset S$  iff

$$\begin{aligned} \|\Delta_{S^c}\|_1 &\leq \sup_{\theta^* \in \mathbb{R}^d} \{\|\theta_S^*\|_1 - \|\Delta_S + \theta_S^*\|_1\} \\ &= \|\Delta_S\|_1 \end{aligned}$$

by the triangle inequality and setting  $\theta_S^* = -\Delta_S$ .