Lecture 2

June 30, 2025

Detour: Martingales

- A filteration $\{\mathcal{F}_k\}_{k \geq 1}$ is a nested sequence of σ-fields $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$.

Detour: Martingales

- A filteration $\{\mathcal{F}_k\}_{k\geq 1}$ is a nested sequence of σ-fields $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$.
- A sequence of random variables $Y_1, Y_2, \ldots$

Detour: Martingales

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

Detour: Martingales

- A filteration $\{\mathcal{F}_k\}_{k\geq 1}$ is a nested sequence of σ-fields $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$.
- A sequence of random variables $Y_1, Y_2, \ldots$
- 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 filteration $\{\mathcal{F}_k\}_{k\geq 1}$ is a nested sequence of σ-fields $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$.
- A sequence of random variables $Y_1, Y_2, \ldots$
- 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 $k \geq 1$.
  - $\mathbb{E}\left(|Y_k|\right) < \infty$.

Detour: Martingales

- A filteration $\{\mathcal{F}_k\}_{k\geq 1}$ is a nested sequence of σ-fields $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$.
- A sequence of random variables $Y_1, Y_2, \ldots$
- 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 $k \geq 1$.
  - $\mathbb{E}\left(|Y_k|\right) < \infty$.
  - $\mathbb{E}(Y_{k+1}|\mathcal{F}_k) = Y_k$ for all $k \geq 1$.

Detour: Martingales

- A filteration $\{\mathcal{F}_k\}_{k\geq 1}$ is a nested sequence of σ-fields $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$.
- A sequence of random variables $Y_1, Y_2, \ldots$
- 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 $k \geq 1$.
  - $\mathbb{E}\left(|Y_k|\right) < \infty$.
  - $\mathbb{E}(Y_{k+1}|\mathcal{F}_k) = Y_k$ for all $k \geq 1$.
- Often $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, in which case we say $\{Y_k\}$ is martingale with respect to $\{X_k\}$.

Detour: Martingales

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

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

Detour: Martingales

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

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

- One of the most general dependence structures in probability.

Detour: Martingales

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

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

- 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

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

Detour: Martingale Examples

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

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

Detour: Martingale Examples

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

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

Detour: Martingale Examples

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

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

Detour: Martingale Examples

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

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

Detour: Martingale Examples

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

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

Detour: Martingale Examples

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

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

- 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

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

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

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

Detour: Martingale Examples

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

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

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

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

Detour: Martingale Examples

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

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

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

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

Detour: Martingale Examples

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

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

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

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

Detour: Martingale Examples

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

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

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

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

- Likelhood 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

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

- Likelhood 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

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

This is a martingale since

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

- Likelhood 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

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

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:

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

- Likelhood 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

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

  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:

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

  Notice that

$$\mathbb{E}(Y_{k+1}|X_1, \ldots, X_k) = \mathbb{E}\left(\mathbb{E}\left(Z \mid X_1, \ldots X_{k+1}\right) \mid X_1, \ldots X_k\right)$$

- Likelhood 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

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

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:

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

Notice that

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

- Likelhood 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

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

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:

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

Notice that

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

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

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

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

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

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

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

$$Y_i \; := \; \mathbb{E}_i\left(Z\right) \; := \; \mathbb{E}\left(Z | X_1, \ldots, X_i\right)$$

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

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

almost surely for all $i = 1, \ldots, n$. 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

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

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 $\mathbb{E}_0\left(Z\right) = \mathbb{E}\left(Z\right)$ and $Y_n = \mathbb{E}_n(Z) = \mathbb{E}(f(X)|X_1, \ldots, X_n) = f(X) = Z$,

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

.

Proof

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

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

- 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 $\mathbb{E}_0\left(Z\right) = \mathbb{E}\left(Z\right)$ and $Y_n = \mathbb{E}_n(Z) = \mathbb{E}(f(X)|X_1, \ldots, X_n) = f(X) = Z$,

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

- 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 $\mathbb{E}_0\left(Z\right) = \mathbb{E}\left(Z\right)$ and $Y_n = \mathbb{E}_n(Z) = \mathbb{E}(f(X)|X_1, \ldots, X_n) = f(X) = Z$,

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

- 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:

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

Proof

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

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

- 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:

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

Proof

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

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

- 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:

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

- Repeating the process, we get

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

- Or

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

and the claim follows.

Bounded difference inequality

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

$$\left| f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \\ \leq L_i \tag{2}$$

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

Bounded difference inequality

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

$$\left|f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)\right| \le L_i \tag{2}$$

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

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

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

Bounded difference inequality

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

$$\left| f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \\ \leq L_i \tag{2}$$

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

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

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

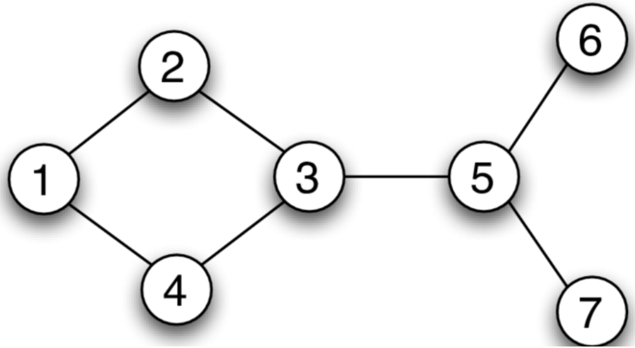
- **Example.** $f(X) = \sum_{i=1}^n X_i$, $X_i \in [a_i, b_i]$.

Clique number of Erdős-Rényi

- Let G be an undirected graph on n nodes.

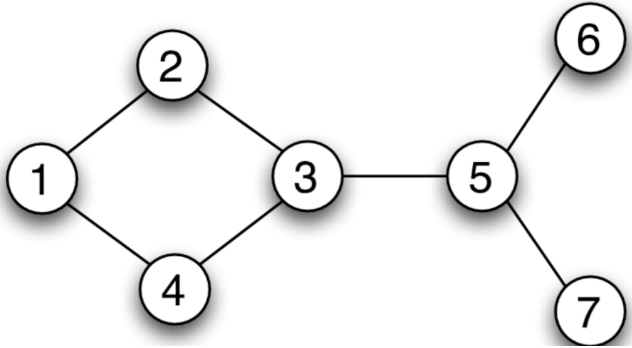
Clique number of Erdős-Rényi

- Let G be an undirected graph on n nodes.



Clique number of Erdős-Rényi

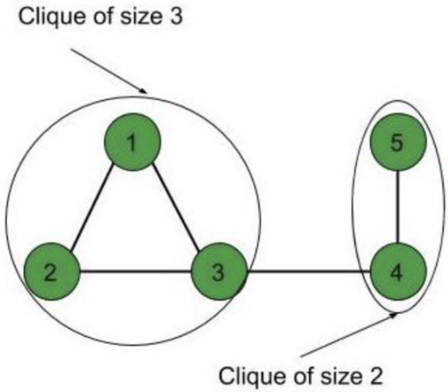
- Let G be an undirected graph on n nodes.



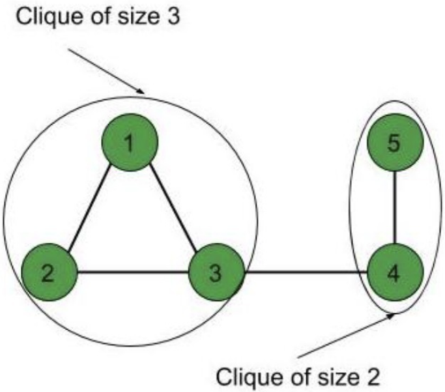
- The edge set is
  $E(G) := \{\{1,2\},\{1,4\},\{2,3\},\{3,4\},\{3,5\},\{5,6\},\{5,7\}\}$.

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lipschitz functions of standard Gaussian vector

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

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

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

Lipschitz functions of standard Gaussian vector

- A function $f : \mathbb{R}^n \to \mathbb{R}$ 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 $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^{-x^2}$ is Lipschitz.
- The function $f : [0,1] \to \mathbb{R}$, $f(x) = \sqrt{x}$ is not Lipschitz.

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

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

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

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

- Notice that it follows that

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

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

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

- Notice that it follows that

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

- 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 $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz w.r.t. the Euclidean norm. Then,

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

- Notice that it follows that

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

- 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 $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz w.r.t. the Euclidean norm. Then,

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

- Notice that it follows that

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

- 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 $\sigma_1(X) \geq \sigma_2(X) \geq \ldots \geq \sigma_k(X)$ be (ordered) singular values of X.

Example: Singular values

- Consider a matrix $X \in \mathbb{R}^{n \times d}$ where $n > d$.
- Let $\sigma_1(X) \geq \sigma_2(X) \geq \ldots \geq \sigma_k(X)$ 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 $\sigma_1(X) \geq \sigma_2(X) \geq \ldots \geq \sigma_k(X)$ 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, $X \to \sigma_1(X)$ is 1-Lipschitz.

Example: Singular values

- Consider a matrix $X \in \mathbb{R}^{n \times d}$ where $n > d$.
- Let $\sigma_1(X) \geq \sigma_2(X) \geq \ldots \geq \sigma_k(X)$ 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, $X \to \sigma_1(X)$ is 1-Lipschitz.
- **Proposition 6.** Let $X \in \mathbb{R}^{n \times d}$ be a random matrix with iid $N(0, 1)$ entries. Then,

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

Example: Singular values

- Consider a matrix $X \in \mathbb{R}^{n \times d}$ where $n > d$.
- Let $\sigma_1(X) \geq \sigma_2(X) \geq \ldots \geq \sigma_k(X)$ 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, $X \to \sigma_1(X)$ is 1-Lipschitz.
- **Proposition 6.** Let $X \in \mathbb{R}^{n \times d}$ be a random matrix with iid $N(0, 1)$ entries. Then,

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

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

Sparsity models

- When $d > n$ there is no hope of estimating θ^*,

Sparsity models

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

Sparsity models

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

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

Sparsity models

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

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

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

Sparsity models

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

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

- Hard sparsity assumption: $s = |S(\theta^*)| << d$.
- Weak sparsity via ℓ_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 θ^*,
- unless we impose a low dimensional assumption on θ^*.
- Support of θ^* (recall $[d] = \{1, \ldots, d\}$)

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

- Hard sparsity assumption: $s = |S(\theta^*)| << d$.
- Weak sparsity via ℓ_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 ℓ_1 ball.

Sparsity models

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

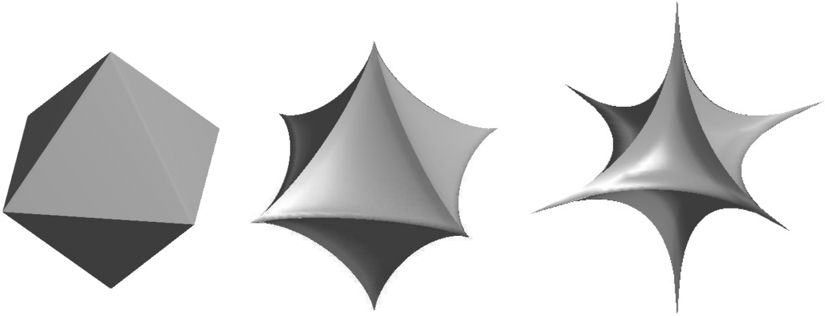
$$\operatorname{supp}(\theta^*) := S(\theta^*) = \{j \in [d] \, : \, \theta_j^* \neq 0\}.$$

- Hard sparsity assumption: $s = |S(\theta^*)| << d$.
- Weak sparsity via ℓ_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 ℓ_1 ball.
- $q = 0$ gives the ℓ_0 ball, same as hard sparsity:

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



(a) (b) (c)

Figure 7.1 Illustrations of the ℓ_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 ℓ_1-ball shown here. (b) For $q = 0.75$, the ball is a non-convex set obtained by collapsing the faces of the ℓ_1-ball towards the origin. (c) For $q = 0.5$, the set becomes more "spiky", and it collapses into the hard sparsity constraint as $q \to 0^+$. 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 θ to an (affine) subspace of $\mathbb{R}^d$, but does not determine θ 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 θ^* 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 θ^* is expected to be “spiky” - that is, concentrated in only a few of its coordinates.
- A model case occurs when θ^* is known to be s-sparse for some $1 \leq s \leq d$, which means that at most s of the coefficients of θ^* 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 θ^*, 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.

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:

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

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

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

Basis pursuit

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

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

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

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

- 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:

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

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

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

- 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

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

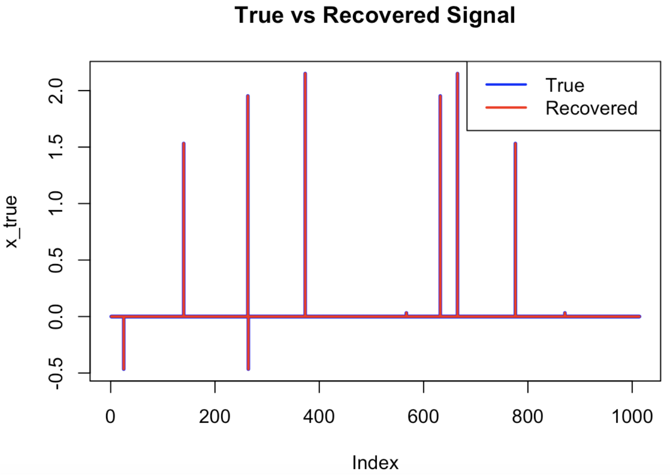
- Global solutions can be obtained very efficiently.

Basis pursuit example

```r
library(CVXR)


set.seed(123)
n <- 1014         # Signal dimension
m <- 64         # Number of measurements
A <- matrix(rnorm(m * n), m, n)  # Measurement matrix



x_true <- rep(0, n)
x_true[sample(1:n, 10)] <- rnorm(5)  # Sparse true signal
plot(x_true)
b <- A %*% x_true


x <- Variable(n)
objective <- Minimize(norm1(x))
constraints <- list(A %*% x == b)
problem <- Problem(objective, constraints)
```

```r
result <- solve(problem)
x_est <- result$getValue(x)


# Plot results
plot(x_true, type = "h", lwd = 3, col = "blue", ylim = range(c(x_true, x_est)),
     main = "True vs Recovered Signal")
lines(x_est, type = "h", col = "red", lwd = 2)
legend("topright", legend = c("True", "Recovered"), col = c("blue", "red"), lwd = 2)

mean((x_est - x_true)^2)
###########################################
```

Basis pursuit example



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

Restricted null space property (RNS)

- Define

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

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 ℓ_1 ball (of radius $\|\theta^*\|_1$) at θ^*:

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

Proof

- Consider the tangent cone to the ℓ_1 ball (of radius $\|\theta^*\|_1$) at θ^*:

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

  i.e., the set of descent directions for the norm ℓ_1 at the point θ^*.

Proof

- Consider the tangent cone to the ℓ_1 ball (of radius $\|\theta^*\|_1$) at θ^*:

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

  i.e., the set of descent directions for the norm ℓ_1 at the point θ^*.
- 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 ℓ_1 ball (of radius $\|\theta^*\|_1$) at θ^*:

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

  i.e., the set of descent directions for the norm ℓ_1 at the point θ^*.
- 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 θ^* if and only if

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

Proof

- Consider the tangent cone to the ℓ_1 ball (of radius $\|\theta^*\|_1$) at θ^*:

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

  i.e., the set of descent directions for the norm ℓ_1 at the point θ^*.
- 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 θ^* if and only if

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

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

Proof

- Consider the tangent cone to the ℓ_1 ball (of radius $\|\theta^*\|_1$) at θ^*:

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

  i.e., the set of descent directions for the norm ℓ_1 at the point θ^*.
- 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 θ^* if and only if

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

  And so the minizer is θ^* if and only if $\mathbb{T}(\theta^*) \cap \ker(X) = \{0\}$.
- It is enough to show that

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

Proof

- Consider the tangent cone to the ℓ_1 ball (of radius $\|\theta^*\|_1$) at θ^*:

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

  i.e., the set of descent directions for the norm ℓ_1 at the point θ^*.
- 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 θ^* if and only if

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

  And so the minizer is θ^* if and only if $\mathbb{T}(\theta^*) \cap \ker(X) = \{0\}$.
- It is enough to show that

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

  since then

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

- It is enough to show that

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

- It is enough to show that

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

- 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

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

- 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

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

- 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

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

- 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

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

- 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

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

- It is enough to show that

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

- 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 θ^* with $\mathrm{supp}(\theta^*) \subset S$ iff

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

- 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 θ^* with $\mathrm{supp}(\theta^*) \subset S$ 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 $\theta^*_S = -\Delta_S$.