Estimation from heavy-tailed data

Stas Minsker

Department of Mathematics, USC

Workshop on Computational Efficiency
and High-Dimensional Robust Statistics

One of the challenges in contemporary statistics and data science is noisy and corrupted data.

One of the challenges in contemporary statistics and data science is noisy and corrupted data.

- Presence of outliers of unknown nature:

  $\Longrightarrow$ requires algorithms that are robust.

One of the challenges in contemporary statistics and data science is noisy and corrupted data.

- Presence of outliers of unknown nature:

  $\Longrightarrow$ requires algorithms that are robust.
- We would like to develop general methods that work under minimal assumptions.

One of the challenges in contemporary statistics and data science is noisy and corrupted data.

- Presence of outliers of unknown nature:

  $\Longrightarrow$ requires algorithms that are robust.
- We would like to develop general methods that work under minimal assumptions.
- A natural way to model outliers is via heavy-tailed distributions.

One of the challenges in contemporary statistics and data science is noisy and corrupted data.

- Presence of outliers of unknown nature:

  $\Longrightarrow$ requires algorithms that are robust.
- We would like to develop general methods that work under minimal assumptions.
- A natural way to model outliers is via heavy-tailed distributions.
- For the purpose of this talk, a random variable X has heavy-tailed distribution if

$$\mathbb{E}|X|^k = \infty$$

  for some $k > 0$ (for example, $k = 2.01$).

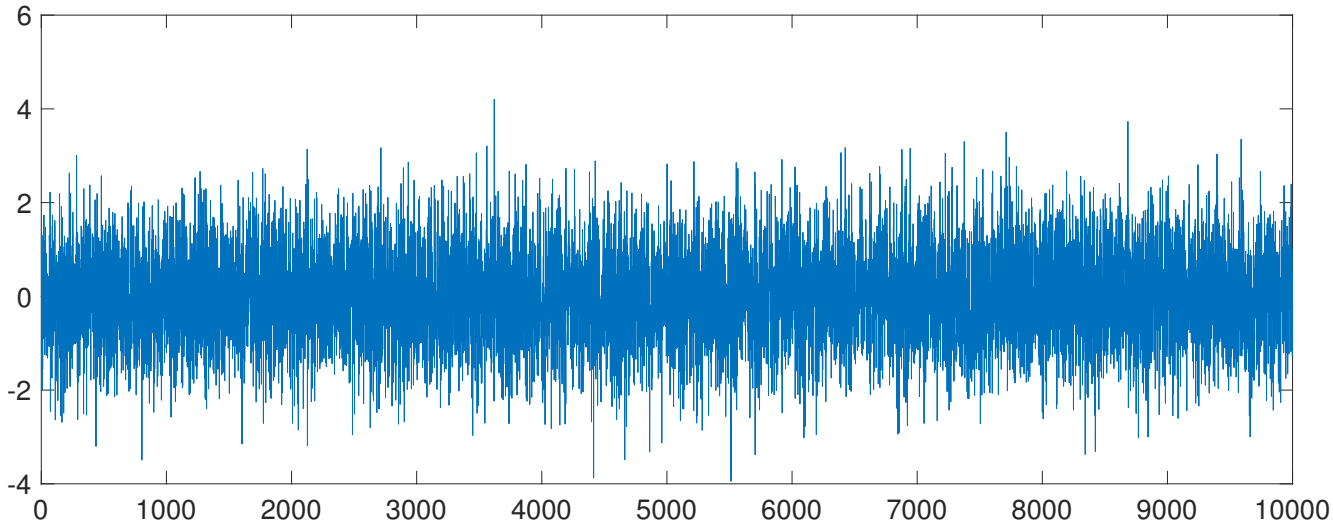


Figure: Standard normal distribution.

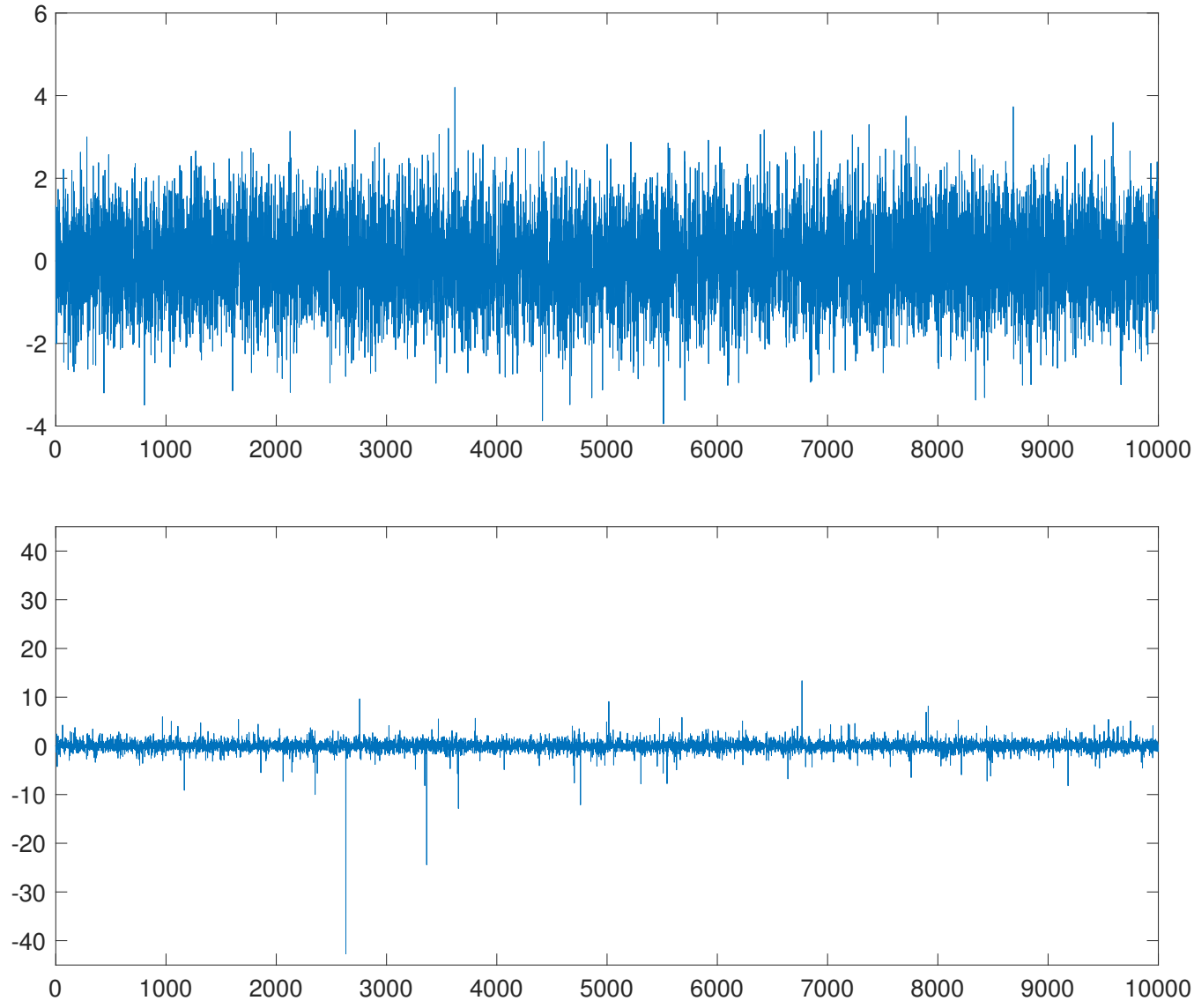


Figure: Student's t-distribution with 3 d.f.

- Main examples for this talk: estimation of the mean and covariance matrix.

Question: how to estimate the mean?

- Assume that $X_1, \ldots, X_N$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$.

Question: how to estimate the mean?

- Assume that $X_1, \ldots, X_N$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$.
- The sample mean $\hat{\mu}_N := \frac{1}{N} \sum_{j=1}^{N} X_j$ satisfies

$$\Pr\left(\left| \hat{\mu}_N - \mu \right| \geq \sigma \sqrt{\frac{2 \log(1/\alpha)}{N}} \right) \leq 2\alpha.$$

Question: how to estimate the mean?

- P. J. Huber (1964): "...This raises a question which could have been asked already by Gauss, but which was, as far as I know, only raised a few years ago (notably by Tukey): what happens if the true distribution deviates slightly from the assumed normal one?"

  What if $X_1, \dots, X_N$ are i.i.d. copies of $X \sim \Pi$ such that

  $$\mathbb{E}X = \mu, \ \mathrm{Var}(X) \leq \sigma^2?$$

  on Π – possibly asymmetric, with heavy tails.

Question: how to estimate the mean?

- P. J. Huber (1964): "...This raises a question which could have been asked already by Gauss, but which was, as far as I know, only raised a few years ago (notably by Tukey): what happens if the true distribution deviates slightly from the assumed normal one?"

  What if $X_1, \ldots, X_N$ are i.i.d. copies of $X \sim \Pi$ such that

  $$\mathbb{E}X = \mu, \ \mathrm{Var}(X) \leq \sigma^2?$$

  on Π – possibly asymmetric, with heavy tails.

- Guarantees for the sample mean $\hat{\mu}_n = \frac{1}{N}\sum_{j=1}^{N} X_j$ are not completely satisfactory:

$$\Pr\left(\left|\hat{\mu}_N - \mu\right| \geq \sigma\sqrt{\frac{(1/\alpha)}{N}}\right) \leq \alpha.$$

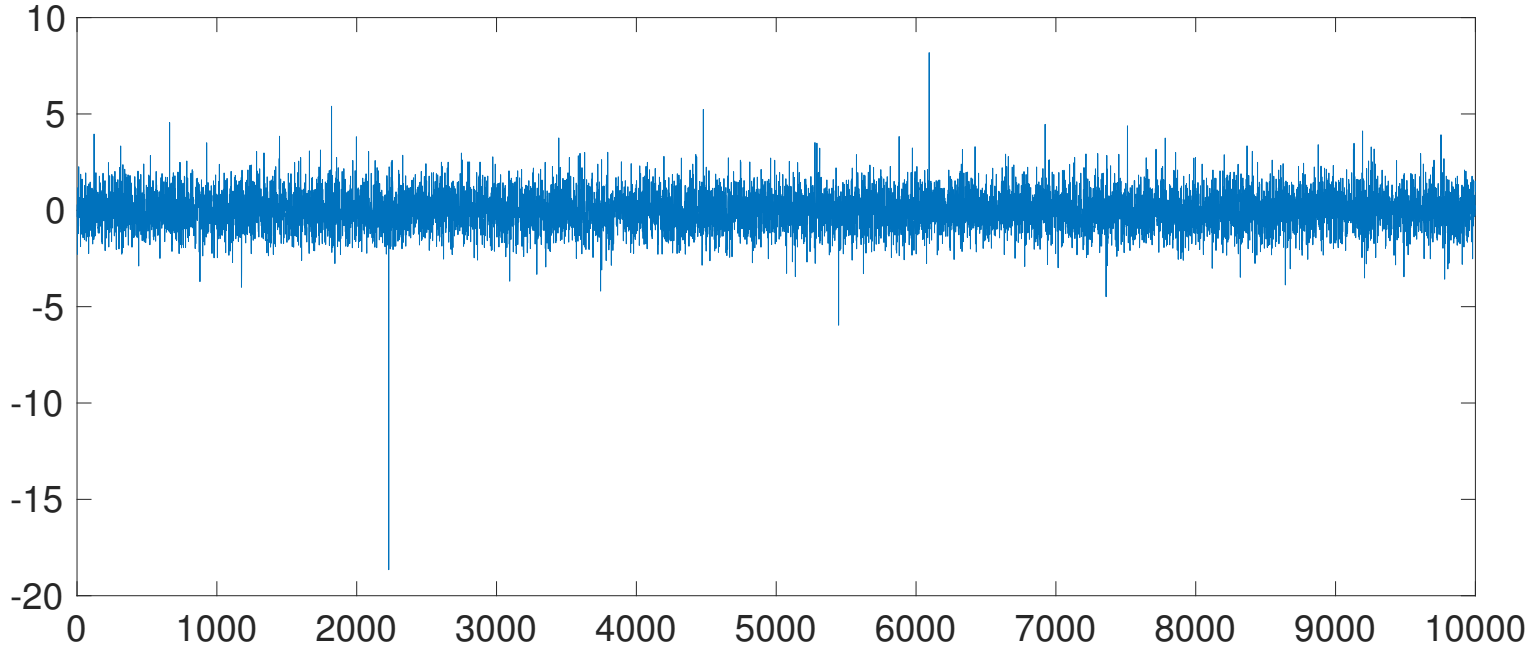


Figure: Rescaled Sample Means of Student's t-distribution with **3 d.f.**

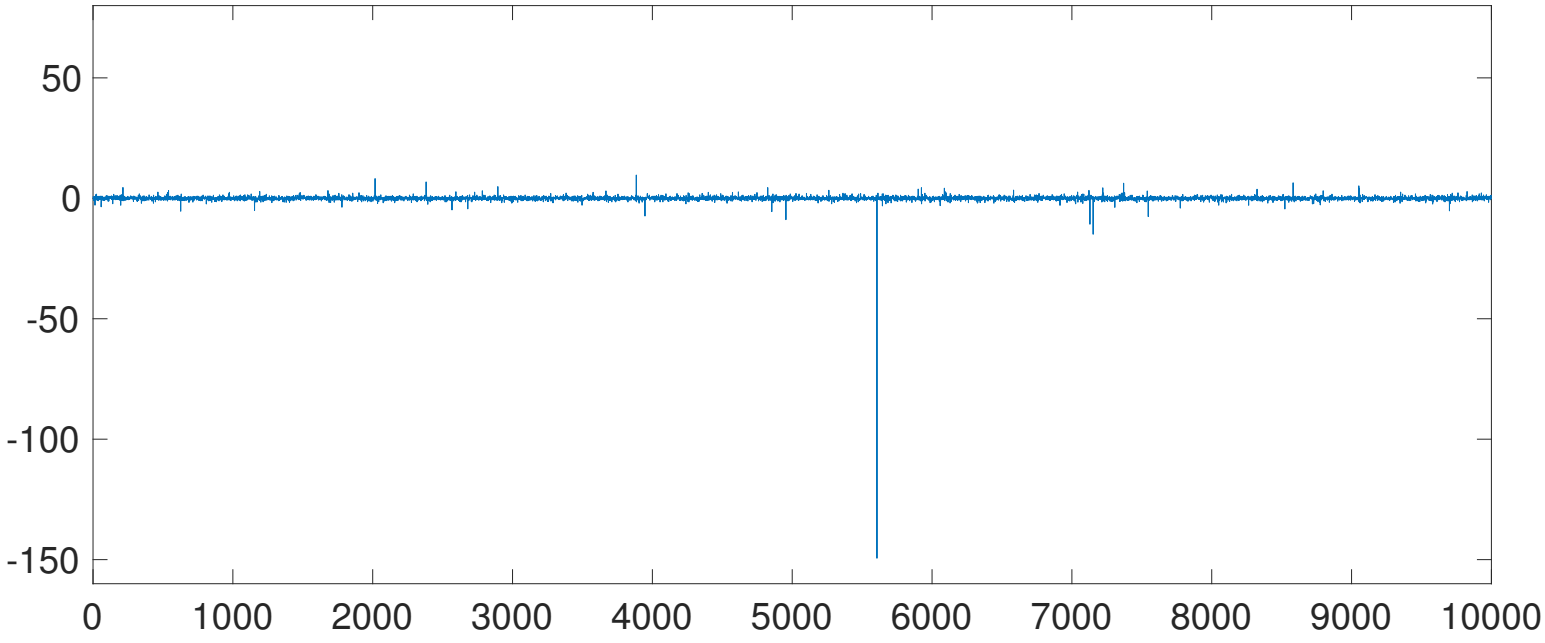


Figure: Rescaled Sample Means of Student's t-distribution with **2.1 d.f.**

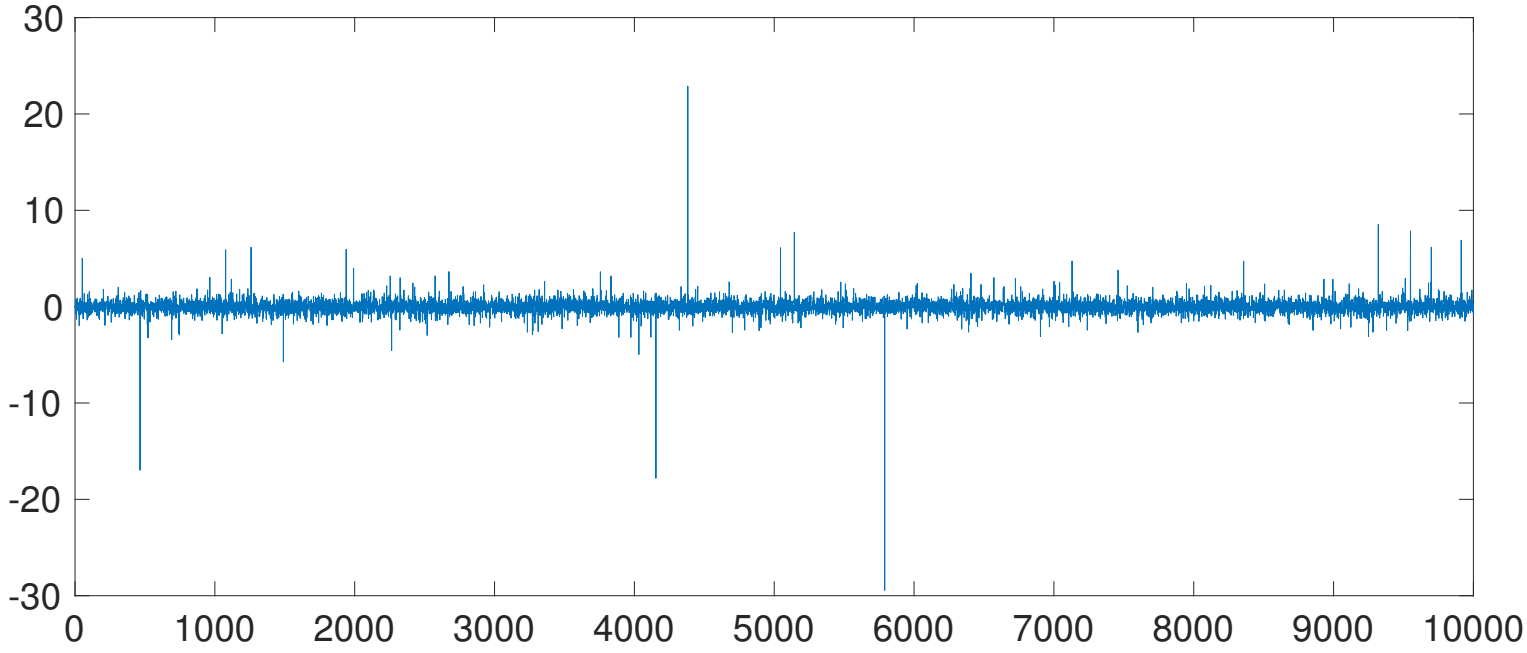


Figure: Rescaled Sample Means of Student's t-distribution with **2.1 d.f.**

Question: how to estimate the mean?

- Median-of-means (MOM) estimator: *[A. Nemirovski, D. Yudin '83; N. Alon, Y. Matias, M. Szegedy '96; R. Oliveira, M. Lerasle '11]*
  Split the sample into $k = \lfloor \log(1/\alpha) \rfloor + 1$ groups $G_1, \dots, G_k$ of size $\simeq N/k$ each:

$$\widehat{\mu}^{(k)} := \text{median}\Big(\underbrace{\overbrace{X_1, \dots, X_{|G_1|}}^{G_1}}_{\bar{\mu}_1 := \frac{1}{|G_1|} \sum_{X_i \in G_1} X_i} \dots\dots \underbrace{\overbrace{X_{N-|G_k|+1}, \dots, X_N}^{G_k}}_{\bar{\mu}_k := \frac{1}{|G_k|} \sum_{X_i \in G_k} X_i} \Big)$$

Question: how to estimate the mean?

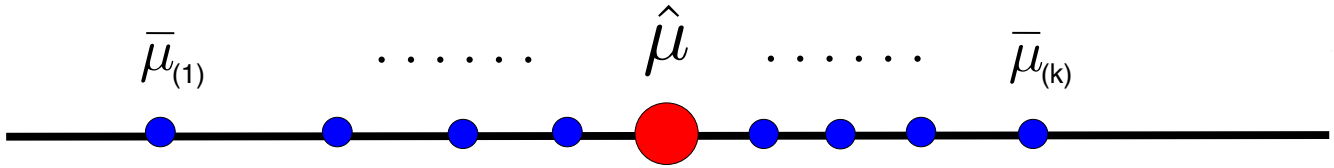
- Median-of-means (MOM) estimator: *[A. Nemirovski, D. Yudin '83; N. Alon, Y. Matias, M. Szegedy '96; R. Oliveira, M. Lerasle '11]*
  Split the sample into $k = \lfloor \log(1/\alpha) \rfloor + 1$ groups $G_1, \ldots, G_k$ of size $\simeq N/k$ each:

$$\widehat{\mu}^{(k)} := \text{median}\Big(\underbrace{\overbrace{X_1, \ldots, X_{|G_1|}}^{G_1}}_{\bar{\mu}_1 := \frac{1}{|G_1|} \sum_{X_i \in G_1} X_i} \; \ldots\ldots \; \underbrace{\overbrace{X_{N-|G_k|+1}, \ldots, X_N}^{G_k}}_{\bar{\mu}_k := \frac{1}{|G_k|} \sum_{X_i \in G_k} X_i} \Big)$$

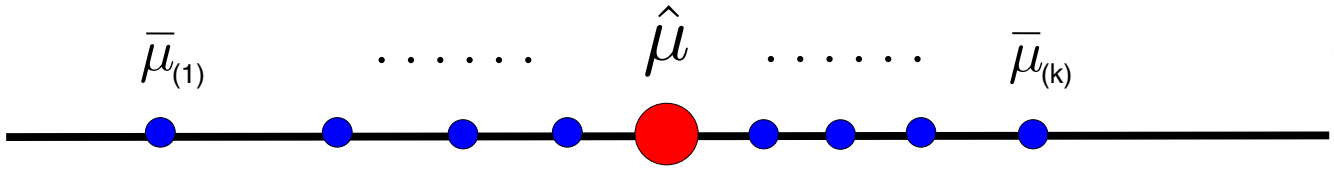
- Claim:

$$\Pr\left(|\widehat{\mu}^{(k)} - \mu| \geq 6.4 \times \sigma \sqrt{\frac{\log(1/\alpha)}{N}} \right) \leq \alpha$$

Proof of the claim ("voting"):

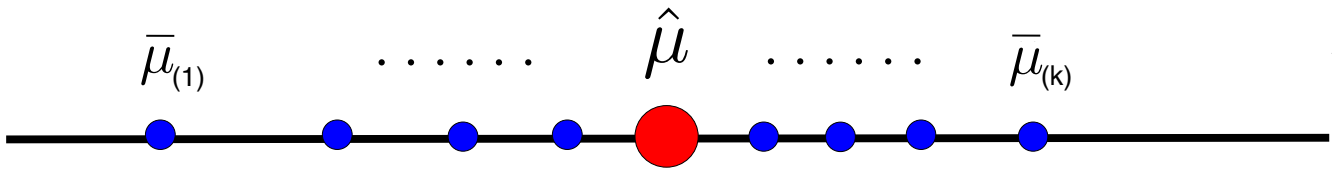


Proof of the claim ("voting"):



$|\hat{\mu}^{(k)} - \mu| \geq s \implies$ at least half of events $\{|\bar{\mu}_j - \mu| \geq s\}$ occur.

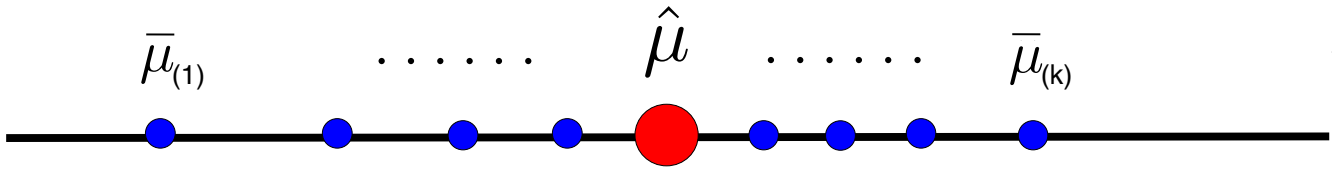
Proof of the claim ("voting"):



$|\hat{\mu}^{(k)} - \mu| \geq s \implies$ at least half of events $\{|\bar{\mu}_j - \mu| \geq s\}$ occur.

Pr (at least half of events $\{|\bar{\mu}_j - \mu| \geq s\}$ occur)

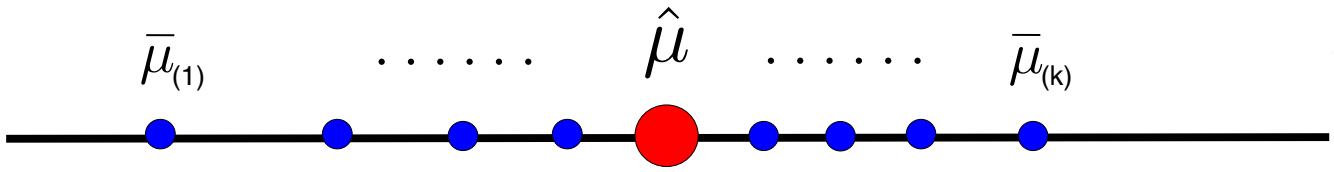
Proof of the claim ("voting"):



$$|\hat{\mu}^{(k)} - \mu| \geq s \implies \text{at least half of events } \{|\bar{\mu}_j - \mu| \geq s\} \text{ occur.}$$

$$\Pr(\text{at least half of events } \{|\bar{\mu}_j - \mu| \geq s\} \text{ occur }) \leq \binom{k}{k/2} (\Pr(|\bar{\mu}_1 - \mu| \geq s))^{k/2}$$

Proof of the claim ("voting"):

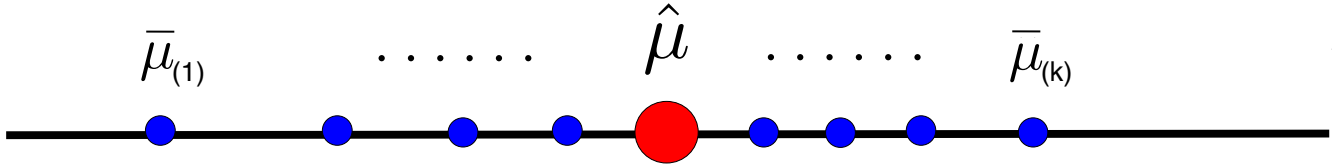


$|\hat{\mu}^{(k)} - \mu| \geq s \implies$ at least half of events $\{|\bar{\mu}_j - \mu| \geq s\}$ occur.

$$\Pr(\text{at least half of events } \{|\bar{\mu}_j - \mu| \geq s\} \text{ occur }) \leq \binom{k}{k/2} (\Pr(|\bar{\mu}_1 - \mu| \geq s))^{k/2}$$

$$\langle \text{Chebyshev's inequality} \rangle \leq (2e)^{k/2} \left(\text{Var}(X) \frac{k}{Ns^2} \right)^{k/2}$$

Proof of the claim ("voting"):



$|\hat{\mu}^{(k)} - \mu| \geq s \implies$ at least half of events $\{|\bar{\mu}_j - \mu| \geq s\}$ occur.

$$\Pr(\text{at least half of events } \{|\bar{\mu}_j - \mu| \geq s\} \text{ occur }) \leq \binom{k}{k/2} (\Pr(|\bar{\mu}_1 - \mu| \geq s))^{k/2}$$

$$\leq (2e)^{k/2} \left(\mathrm{Var}(X) \frac{k}{Ns^2} \right)^{k/2} \leq e^{-k}$$

whenever $s \geq (2e^3)^{1/2} \sigma \sqrt{\frac{k}{N}}$.

Proof of the claim ("voting"):

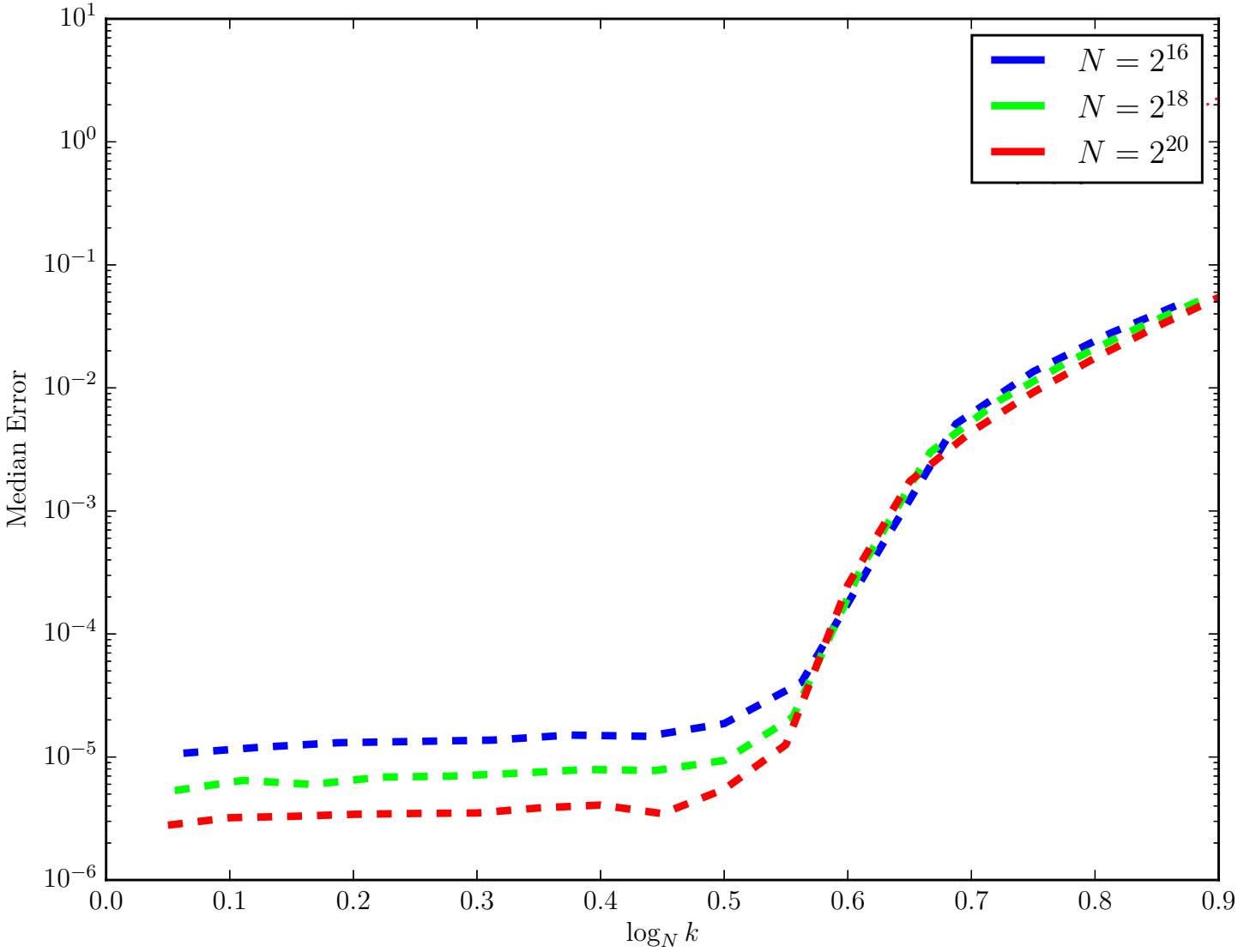
$$|\hat{\mu}^{(k)} - \mu| \geq s \implies \text{ at least half of events } \{|\bar{\mu}_j - \mu| \geq s\} \text{ occur.}$$

$$\begin{aligned}
&\Pr(\text{at least half of events } \{|\bar{\mu}_j - \mu| \geq s\} \text{ occur }) \\
&\leq \binom{k}{k/2} (\Pr(|\bar{\mu}_1 - \mu| \geq s))^{k/2} \\
&\leq (2e)^{k/2} \left(\mathrm{Var}(X) \frac{k}{Ns^2} \right)^{k/2} \leq e^{-k}
\end{aligned}$$

$$\text{whenever } s \geq (2e^3)^{1/2} \sigma \sqrt{\tfrac{k}{N}}.$$

- Since $k = \lfloor \log(1/\alpha) \rfloor + 1$, result holds for an "absolute constant" $(2e^3)^{1/2} \leq 6.4$.

Perfomance as *k* changes



Result so far:

$$k = \lfloor \log(1/\alpha) \rfloor + 1$$

$$\Pr\left(|\widehat{\mu}^{(k)} - \mu| \geq 6.4\sigma\sqrt{\frac{k}{N}} \right) \leq \alpha$$

- Need to recalculate the estimator for different values of α.

Result so far:

$$k = \lfloor \log(1/\alpha) \rfloor + 1$$

$$\Pr\left(|\widehat{\mu}^{(k)} - \mu| \geq 6.4\sigma\sqrt{\frac{k}{N}} \right) \leq \alpha$$

- Need to recalculate the estimator for different values of α.
- Estimator depends on the random partition of the index set.

Result so far:

$$k = \lfloor \log(1/\alpha) \rfloor + 1$$

$$\Pr\left(|\widehat{\mu}^{(k)} - \mu| \geq 6.4\sigma\sqrt{\frac{k}{N}} \right) \leq \alpha$$

- Need to recalculate the estimator for different values of α.
- Estimator depends on the random partition of the index set.
- How to choose k? Typically, want k as large as possible.

Result so far:

$$k = \lfloor \log(1/\alpha) \rfloor + 1$$

$$\Pr\left(|\widehat{\mu}^{(k)} - \mu| \geq 6.4\sigma\sqrt{\frac{k}{N}} \right) \leq \alpha$$

- Need to recalculate the estimator for different values of α.
- Estimator depends on the random partition of the index set.
- How to choose k? Typically, want k as large as possible.
- Limiting distribution: what happens when $k, N \to \infty$?

Result so far:

$$k = \lfloor \log(1/\alpha) \rfloor + 1$$

$$\Pr\left(|\widehat{\mu}^{(k)} - \mu| \geq 6.4\sigma \sqrt{\frac{k}{N}} \right) \leq \alpha$$

- Need to recalculate the estimator for different values of α.
- Estimator depends on the random partition of the index set.
- How to choose k? Typically, want k as large as possible.
- Limiting distribution: what happens when $k, N \to \infty$?
- Confidence intervals?

- Idea: what if the distribution is symmetric?

- Idea: what if the distribution is symmetric?
- The mean and the median coincide $\Longrightarrow$ many robust estimators to choose from.

- Idea: what if the distribution is symmetric?
- The mean and the median coincide $\Longrightarrow$ many robust estimators to choose from.
- If the distribution is not symmetric, create a new sample that is "almost symmetric" and has the same mean.

- Idea: what if the distribution is symmetric?
- The mean and the median coincide $\implies$ many robust estimators to choose from.
- If the distribution is not symmetric, create a new sample that is "almost symmetric" and has the same mean.
- Works in other problems, such as MLE.

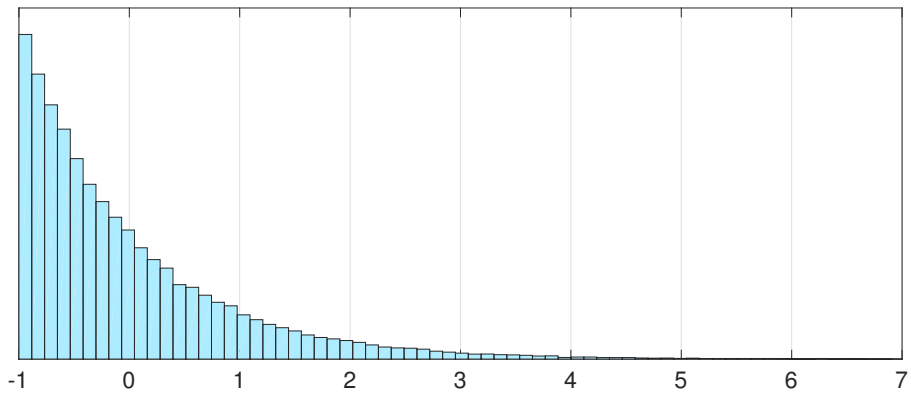


Figure: Centered exponential distribution

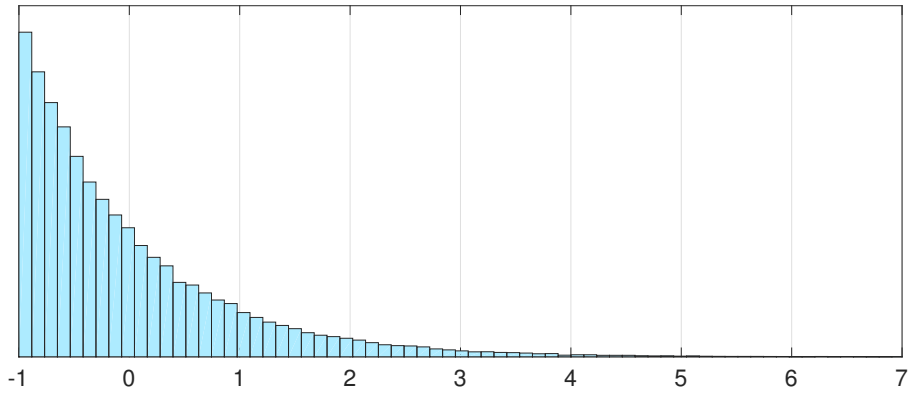


Figure: Centered exponential distribution

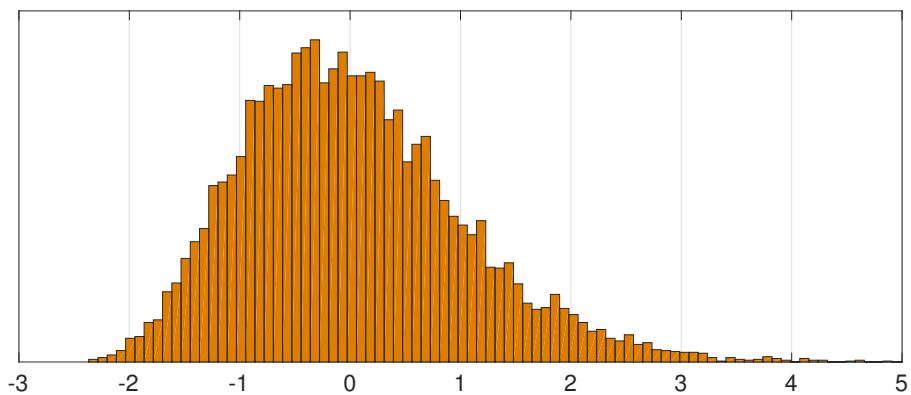


Figure: Rescaled sample means with $n = 10$.

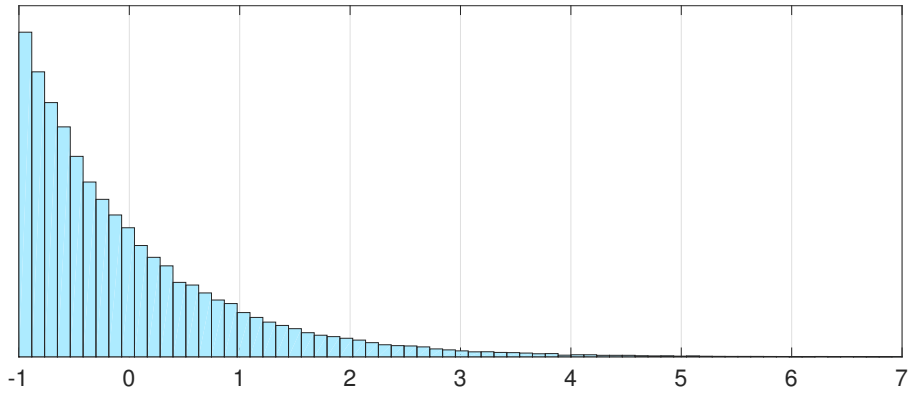


Figure: Centered exponential distribution

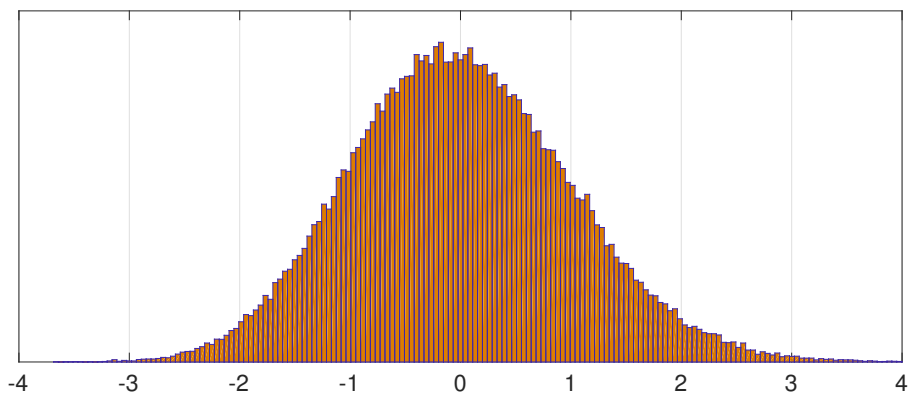


Figure: Rescaled sample means with $n = 100$.

- Idea for the Theorem: rates of convergence to normal distribution are "traded" for tight concentration.

- Idea for the Theorem: rates of convergence to normal distribution are "traded" for tight concentration.
- $X_1, \ldots, X_N$ - independent observations from $P_1, \ldots, P_N$.

- Idea for the Theorem: rates of convergence to normal distribution are "traded" for tight concentration.
- $X_1, \ldots, X_N$ - independent observations from $P_1, \ldots, P_N$.
  θ_* – shared parameter (e.g., mean).

- Idea for the Theorem: rates of convergence to normal distribution are "traded" for tight concentration.
- $X_1, \ldots, X_N$ - independent observations from $P_1, \ldots, P_N$.
  θ_* – shared parameter (e.g., mean).
- Data are split into k disjoint groups $G_1, \ldots, G_k$ of cardinalities $n_1, \ldots, n_k$.

- Idea for the Theorem: rates of convergence to normal distribution are "traded" for tight concentration.
- $X_1, \ldots, X_N$ - independent observations from $P_1, \ldots, P_N$.
  θ_* – shared parameter (e.g., mean).
- Data are split into k disjoint groups $G_1, \ldots, G_k$ of cardinalities $n_1, \ldots, n_k$.
- $\bar{\theta}_j = \bar{\theta}_j(X_i,\ i \in G_j)$.

- Idea for the Theorem: rates of convergence to normal distribution are "traded" for tight concentration.
- $X_1, \ldots, X_N$ - independent observations from $P_1, \ldots, P_N$.
  θ_* – shared parameter (e.g., mean).
- Data are split into k disjoint groups $G_1, \ldots, G_k$ of cardinalities $n_1, \ldots, n_k$.
- $\bar{\theta}_j = \bar{\theta}_j(X_i,\ i \in G_j)$.

$$\widehat{\theta}_\rho^{(k)} := \operatorname*{argmin}_{z \in \mathbb{R}} \sum_{j=1}^{k} \rho\left(z - \bar{\theta}_j\right)$$

- Idea for the Theorem: rates of convergence to normal distribution are "traded" for tight concentration.
- $X_1, \ldots, X_N$ - independent observations from $P_1, \ldots, P_N$.
  θ_* – shared parameter (e.g., mean).
- Data are split into k disjoint groups $G_1, \ldots, G_k$ of cardinalities $n_1, \ldots, n_k$.
- $\bar{\theta}_j = \bar{\theta}_j(X_i,\ i \in G_j)$.

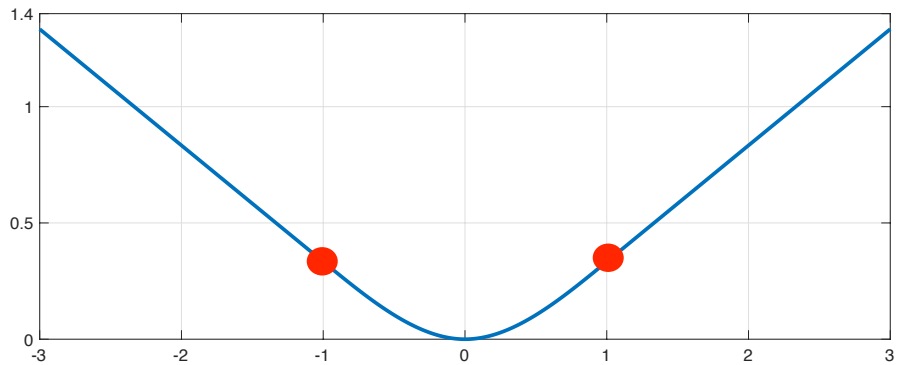
$$\widehat{\theta}_\rho^{(k)} := \operatorname*{argmin}_{z \in \mathbb{R}} \sum_{j=1}^{k} \rho\left(z - \bar{\theta}_j\right)$$

- $\rho(x) = |x|$, or

- Idea for the Theorem: rates of convergence to normal distribution are "traded" for tight concentration.
- $X_1, \ldots, X_N$ - independent observations from $P_1, \ldots, P_N$.
  θ_* – shared parameter (e.g., mean).
- Data are split into k disjoint groups $G_1, \ldots, G_k$ of cardinalities $n_1, \ldots, n_k$.
- $\bar{\theta}_j = \bar{\theta}_j(X_i,\ i \in G_j)$.

$$\widehat{\theta}_\rho^{(k)} := \operatorname*{argmin}_{z \in \mathbb{R}} \sum_{j=1}^{k} \rho\left(z - \bar{\theta}_j\right)$$

- $\rho(x) = |x|$, or
- $\rho(x) =$ Huber's loss



- $\Phi(t)$ – distribution function of $N(0,1)$.

- $\Phi(t)$ – distribution function of $N(0,1)$.
- Assumption: $\bar{\theta}_1, \ldots, \bar{\theta}_k$ are asymptotically normal.

- $\Phi(t)$ – distribution function of $N(0,1)$.
- Assumption: $\bar{\theta}_1, \ldots, \bar{\theta}_k$ are asymptotically normal.
- For each $j = 1, \ldots, k$ (number of estimators),

- $\Phi(t)$ – distribution function of $N(0,1)$.
- Assumption: $\bar{\theta}_1, \ldots, \bar{\theta}_k$ are asymptotically normal.
- For each $j = 1, \ldots, k$,
  there exist a sequence $\underbrace{\{\sigma_n^{(j)}\}_{n\in\mathbb{N}}}_{\text{standard deviations}}$ such that

- $\Phi(t)$ – distribution function of $N(0,1)$.
- Assumption: $\bar{\theta}_1, \ldots, \bar{\theta}_k$ are asymptotically normal.
- For each $j = 1, \ldots, k$,
  there exist a sequence $\{\sigma_n^{(j)}\}_{n \in \mathbb{N}}$ such that

$$g_j(n_j) := \underbrace{\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\bar{\theta}_j - \theta_*}{\sigma_{n_j}^{(j)}} \leq t \right) - \Phi(t) \right| \to 0}_{\text{Central Limit Theorem}} \text{ as } n_j \to \infty.$$

- $\Phi(t)$ – distribution function of $N(0,1)$.
- Assumption: $\bar{\theta}_1, \ldots, \bar{\theta}_k$ are asymptotically normal.
- For each $j = 1, \ldots, k$,
  there exist a sequence $\{\sigma_n^{(j)}\}_{n \in \mathbb{N}}$ such that

$$g_j(n_j) := \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\bar{\theta}_j - \theta_*}{\sigma_{n_j}^{(j)}} \leq t \right) - \Phi(t) \right| \to 0 \text{ as } n_j \to \infty.$$

- Examples: Berry-Esseen theorem (mean), smooth likelihood (MLE).

Results for MOM (i.e. $\rho(x) = |x|$)

$$g_j(n_j) = \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\bar{\theta}_j - \theta_*}{\sigma_{n_j}^{(j)}} \leq t \right) - \Phi(t) \right|$$

Results for MOM (i.e. $\rho(x) = |x|$)

$$g_j(n_j) = \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\bar{\theta}_j - \theta_*}{\sigma_{n_j}^{(j)}} \leq t \right) - \Phi(t) \right|$$

$$H_k = \left(\frac{1}{k} \sum_{j=1}^{k} \frac{1}{\sigma_{n_j}^{(j)}} \right)^{-1} - \text{Harmonic mean}$$

Results for MOM (i.e. $\rho(x) = |x|$)

$$g_j(n_j) = \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\bar{\theta}_j - \theta_*}{\sigma_{n_j}^{(j)}} \leq t \right) - \Phi(t) \right|$$

$$H_k = \left(\frac{1}{k} \sum_{j=1}^{k} \frac{1}{\sigma_{n_j}^{(j)}} \right)^{-1} \quad - \text{Harmonic mean}$$

$$\alpha_j = \frac{H_k}{\sigma_{n_j}^{(j)}}, \; j = 1, \ldots, k \quad \text{(constants)}$$

Results for MOM (i.e. $\rho(x) = |x|$)

$$g_j(n_j) = \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\bar{\theta}_j - \theta_*}{\sigma_{n_j}^{(j)}} \leq t \right) - \Phi(t) \right|$$

$$H_k = \left(\frac{1}{k} \sum_{j=1}^{k} \frac{1}{\sigma_{n_j}^{(j)}} \right)^{-1} \quad - \text{Harmonic mean}$$

$$\alpha_j = \frac{H_k}{\sigma_{n_j}^{(j)}}, \; j = 1, \ldots, k \quad \text{(constants)}$$

Theorem [M., 2018]

For all $s > 0$ such that $\frac{1}{k} \sum_{i=1}^{k} \left(g_i(n_i) + \sqrt{\frac{s}{k}} \right) \cdot \max_{j=1,\ldots,k} \alpha_j \leq 0.33$ [basically, $s \lesssim k$],

Results for MOM (i.e. $\rho(x) = |x|$)

$$g_j(n_j) = \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\bar{\theta}_j - \theta_*}{\sigma_{n_j}^{(j)}} \le t \right) - \Phi(t) \right|$$

$$H_k = \left(\frac{1}{k} \sum_{j=1}^{k} \frac{1}{\sigma_{n_j}^{(j)}} \right)^{-1} \quad - \text{Harmonic mean}$$

$$\alpha_j = \frac{H_k}{\sigma_{n_j}^{(j)}}, \; j = 1, \ldots, k \quad \text{(constants)}$$

Theorem [M., 2018]

For all $s > 0$ such that $\frac{1}{k} \sum_{i=1}^{k} \left(g_i(n_i) + \sqrt{\frac{s}{k}} \right) \cdot \max_{j=1,\ldots,k} \alpha_j \le 0.33$,

$$\left| \widehat{\theta}^{(k)} - \theta_* \right| \le 3 H_k \cdot \frac{1}{k} \sum_{j=1}^{k} \left(g_j(n_j) + \sqrt{\frac{s}{k}} \right)$$

with probability at least $1 - 4e^{-2s}$.

Corollary

$X_1, \ldots, X_N$ – i.i.d. such that $\mathbb{E}X_1 = \theta_*$, $\mathrm{Var}(X_1) = \sigma^2$, $\mathbb{E}|X_1 - \theta_*|^{2+\delta} < \infty$ for $\delta \in (0, 1]$.

Corollary

$X_1, \ldots, X_N$ – i.i.d. such that $\mathbb{E}X_1 = \theta_*$, $\mathrm{Var}(X_1) = \sigma^2$, $\mathbb{E}|X_1 - \theta_*|^{2+\delta} < \infty$ for $\delta \in (0, 1]$. Then $\exists$ absolute constants $c_1, c_2 > 0$ such that for all $s > 0$ and k satisfying

$$\frac{\mathbb{E}|X - \theta_*|^{2+\delta}}{\sigma^{2+\delta} n^{\delta/2}} + \sqrt{\frac{s}{k}} \leq c_1,$$

Corollary

$X_1, \ldots, X_N$ – i.i.d. such that $\mathbb{E}X_1 = \theta_*$, $\mathrm{Var}(X_1) = \sigma^2$, $\mathbb{E}|X_1 - \theta_*|^{2+\delta} < \infty$ for $\delta \in (0, 1]$. Then $\exists$ absolute constants $c_1, c_2 > 0$ such that for all $s > 0$ and k satisfying

$$\frac{\mathbb{E}|X - \theta_*|^{2+\delta}}{\sigma^{2+\delta} n^{\delta/2}} + \sqrt{\frac{s}{k}} \leq c_1,$$

the inequality

$$|\widehat{\theta}^{(k)} - \theta_*| \leq c_2 \sigma \left(\frac{\mathbb{E}\,|X - \theta_*|^{2+\delta} / \sigma^{2+\delta}}{n^{\frac{1+\delta}{2}}} + \sqrt{\frac{s}{N}} \right) \qquad \left[n = \lfloor N/k \rfloor \right]$$

holds with probability at least $1 - 4e^{-2s}$.

Corollary

$X_1, \ldots, X_N$ – i.i.d. such that $\mathbb{E}X_1 = \theta_*$, $\mathrm{Var}(X_1) = \sigma^2$, $\mathbb{E}|X_1 - \theta_*|^{2+\delta} < \infty$ for $\delta \in (0, 1]$. Then $\exists$ absolute constants $c_1, c_2 > 0$ such that for all $s > 0$ and k satisfying

$$\frac{\mathbb{E}|X - \theta_*|^{2+\delta}}{\sigma^{2+\delta} n^{\delta/2}} + \sqrt{\frac{s}{k}} \leq c_1,$$

the inequality

$$|\widehat{\theta}^{(k)} - \theta_*| \leq c_2 \sigma \left(\underbrace{\frac{\mathbb{E}\,|X - \theta_*|^{2+\delta} / \sigma^{2+\delta}}{n^{\frac{1+\delta}{2}}}}_{\text{"bias"}} + \sqrt{\frac{s}{N}} \right)$$

holds with probability at least $1 - 4e^{-2s}$.

"Bias" is of smaller order whenever $k \lesssim N^{\delta/(1+\delta)}$.

Idea of the proof (i.i.d. case, equal group sizes)

- $n = \lfloor N/k \rfloor$, $W_j = \sqrt{n}\frac{\bar{\theta}_j - \theta_*}{\sigma}$, $j = 1, \ldots, k$.

Idea of the proof (i.i.d. case, equal group sizes)

- $n = \lfloor N/k \rfloor$, $W_j = \sqrt{n}\frac{\bar{\theta}_j - \theta_*}{\sigma}$, $j = 1, \ldots, k$.
- $\Phi^{(n)}$ – the d.f. of W_1, $\widehat{\Phi}_k$ – the empirical d.f. of $W_1, \ldots, W_k$.

Idea of the proof (i.i.d. case, equal group sizes)

- $n = \lfloor N/k \rfloor$, $W_j = \sqrt{n}\frac{\bar{\theta}_j - \theta_*}{\sigma}$, $j = 1, \ldots, k$.
- $\Phi^{(n)}$ – the d.f. of W_1, $\widehat{\Phi}_k$ – the empirical d.f. of $W_1, \ldots, W_k$.
- Need: $z_1 \geq z_2$ such that

$$\widehat{\Phi}_k(z_1) \geq \frac{1}{2} \text{ and } \widehat{\Phi}_k(z_2) \leq \frac{1}{2} \implies \text{median}(W_1, \ldots, W_k) \in [z_2, z_1]$$

Idea of the proof (i.i.d. case, equal group sizes)

- $n = \lfloor N/k \rfloor$, $W_j = \sqrt{n}\frac{\bar{\theta}_j - \theta_*}{\sigma}$, $j = 1, \ldots, k$.
- $\Phi^{(n)}$ – the d.f. of W_1, $\widehat{\Phi}_k$ – the empirical d.f. of $W_1, \ldots, W_k$.
- Need: $z_1 \geq z_2$ such that

$$\widehat{\Phi}_k(z_1) \geq \frac{1}{2} \text{ and } \widehat{\Phi}_k(z_2) \leq \frac{1}{2} \implies \text{median}(W_1, \ldots, W_k) \in [z_2, z_1]$$

- $\Phi(z)$ – c.d.f of standard normal, then

$$\widehat{\Phi}_k(z_1) = \Phi(z_1) + \widehat{\Phi}_k(z_1) - \Phi^{(n)}(z_1) + \Phi^{(n)}(z_1) - \Phi(z_1)$$

Idea of the proof (i.i.d. case, equal group sizes)

- $n = \lfloor N/k \rfloor$, $W_j = \sqrt{n}\frac{\bar{\theta}_j - \theta_*}{\sigma}$, $j = 1, \ldots, k$.
- $\Phi^{(n)}$ – the d.f. of W_1, $\widehat{\Phi}_k$ – the empirical d.f. of $W_1, \ldots, W_k$.
- Need: $z_1 \geq z_2$ such that

$$\widehat{\Phi}_k(z_1) \geq \frac{1}{2} \text{ and } \widehat{\Phi}_k(z_2) \leq \frac{1}{2} \implies \text{median}(W_1, \ldots, W_k) \in [z_2, z_1]$$

- $\Phi(z)$ – c.d.f of standard normal, then

$$\widehat{\Phi}_k(z_1) \geq \Phi(z_1) - |\widehat{\Phi}_k(z_1) - \Phi^{(n)}(z_1)| - |\Phi^{(n)}(z_1) - \Phi(z_1)|$$

Idea of the proof (i.i.d. case, equal group sizes)

- $n = \lfloor N/k \rfloor$, $W_j = \sqrt{n}\frac{\bar{\theta}_j - \theta_*}{\sigma}$, $j = 1, \ldots, k$.
- $\Phi^{(n)}$ – the d.f. of W_1, $\widehat{\Phi}_k$ – the empirical d.f. of $W_1, \ldots, W_k$.
- Need: $z_1 \geq z_2$ such that

$$\widehat{\Phi}_k(z_1) \geq \frac{1}{2} \text{ and } \widehat{\Phi}_k(z_2) \leq \frac{1}{2} \implies \text{median}(W_1, \ldots, W_k) \in [z_2, z_1]$$

- $\Phi(z)$ – c.d.f of standard normal, then

$$\widehat{\Phi}_k(z_1) \geq \Phi(z_1) - \underbrace{|\widehat{\Phi}_k(z_1) - \Phi^{(n)}(z_1)|}_{\text{bounded difference}} - \underbrace{|\Phi^{(n)}(z_1) - \Phi(z_1)|}_{\text{Berry-Esseen}}$$

Idea of the proof (i.i.d. case, equal group sizes)

- $n = \lfloor N/k \rfloor$, $W_j = \sqrt{n}\frac{\bar{\theta}_j - \theta_*}{\sigma}$, $j = 1, \ldots, k$.
- $\Phi^{(n)}$ – the d.f. of W_1, $\widehat{\Phi}_k$ – the empirical d.f. of $W_1, \ldots, W_k$.
- Need: $z_1 \geq z_2$ such that

$$\widehat{\Phi}_k(z_1) \geq \frac{1}{2} \text{ and } \widehat{\Phi}_k(z_2) \leq \frac{1}{2} \implies \text{median}(W_1, \ldots, W_k) \in [z_2, z_1]$$

- $\Phi(z)$ – c.d.f of standard normal, then

$$\widehat{\Phi}_k(z_1) \geq \Phi(z_1) - \underbrace{|\widehat{\Phi}_k(z_1) - \Phi^{(n)}(z_1)|}_{\text{bounded difference}} - \underbrace{|\Phi^{(n)}(z_1) - \Phi(z_1)|}_{\text{Berry-Esseen}}$$

- Enough to find z_1 such that

$$\Phi(z_1) \geq \frac{1}{2} + \sqrt{\frac{s}{k}} + 0.5\frac{\mathbb{E}|X - \mathbb{E}X|^3}{\sigma^3\sqrt{n}}.$$

Idea of the proof (i.i.d. case, equal group sizes)

- $n = \lfloor N/k \rfloor$, $W_j = \sqrt{n}\frac{\bar{\theta}_j - \theta_*}{\sigma}$, $j = 1, \ldots, k$.
- $\Phi^{(n)}$ – the d.f. of W_1, $\widehat{\Phi}_k$ – the empirical d.f. of $W_1, \ldots, W_k$.
- Need: $z_1 \geq z_2$ such that

$$\widehat{\Phi}_k(z_1) \geq \frac{1}{2} \text{ and } \widehat{\Phi}_k(z_2) \leq \frac{1}{2} \implies \text{median}(W_1, \ldots, W_k) \in [z_2, z_1]$$

- $\Phi(z)$ – c.d.f of standard normal, then

$$\widehat{\Phi}_k(z_1) \geq \Phi(z_1) - \underbrace{|\widehat{\Phi}_k(z_1) - \Phi^{(n)}(z_1)|}_{\text{bounded difference}} - \underbrace{|\Phi^{(n)}(z_1) - \Phi(z_1)|}_{\text{Berry-Esseen}}$$

- Enough to find z_1 such that

$$\Phi(z_1) \geq \frac{1}{2} + \sqrt{\frac{s}{k}} + 0.5\frac{\mathbb{E}|X - \mathbb{E}X|^3}{\sigma^3\sqrt{n}}.$$

$$\implies z_1 \simeq \sqrt{\tfrac{s}{k}} + 0.5\tfrac{\mathbb{E}|X-\mathbb{E}X|^3}{\sigma^3\sqrt{n}}.$$

Eliminating dependence on the partition

- $n = \lfloor N/k \rfloor$

Eliminating dependence on the partition

- $n = \lfloor N/k \rfloor$
- $\mathcal{A}_N^{(n)} = \{J \subset \{1, \ldots, N\} : \ |J| = n\}$

Eliminating dependence on the partition

- $n = \lfloor N/k \rfloor$
- $\mathcal{A}_N^{(n)} = \{J \subset \{1, \ldots, N\} : \ |J| = n\}$
- $\bar{\theta}_J = \bar{\theta}(X_j, \ j \in J)$

Eliminating dependence on the partition

- $n = \lfloor N/k \rfloor$
- $\mathcal{A}_N^{(n)} = \{J \subset \{1, \ldots, N\} : \ |J| = n\}$
- $\bar{\theta}_J = \bar{\theta}(X_j, \ j \in J)$
-

$$\widetilde{\theta}_\rho^{(k)} := \operatorname*{argmin}_{z \in \mathbb{R}} \sum_{J \in \mathcal{A}_N^{(n)}} \rho\left(z - \bar{\theta}_J\right)$$

Eliminating dependence on the partition

- $n = \lfloor N/k \rfloor$
- $\mathcal{A}_N^{(n)} = \{J \subset \{1, \ldots, N\} : \ |J| = n\}$
- $\bar{\theta}_J = \bar{\theta}(X_j, \ j \in J)$
- $$\widetilde{\theta}_\rho^{(k)} := \operatorname*{argmin}_{z \in \mathbb{R}} \sum_{J \in \mathcal{A}_N^{(n)}} \rho\left(z - \bar{\theta}_J\right)$$
- Same (exactly) deviation guarantees as before.

Central Limit Theorem

- Question: what happens when $k, N \to \infty$?

Central Limit Theorem

- Question: what happens when $k, N \to \infty$?
- $X_1, \ldots, X_N$ - independent from $P_1, \ldots, P_N$, θ_* – shared parameter (e.g., mean).

Central Limit Theorem

- Question: what happens when $k, N \to \infty$?
- $X_1, \ldots, X_N$ - independent from $P_1, \ldots, P_N$, θ_* – shared parameter (e.g., mean).
- $\bar{\theta}_1, \ldots, \bar{\theta}_k$ are asymptotically normal, $g_j(n_j)$ control approximation rate.

Central Limit Theorem

- Question: what happens when $k, N \to \infty$?
- $X_1, \ldots, X_N$ - independent from $P_1, \ldots, P_N$, θ_* – shared parameter (e.g., mean).
- $\bar{\theta}_1, \ldots, \bar{\theta}_k$ are asymptotically normal, $g_j(n_j)$ control approximation rate.
- $\max_{j=1,\ldots,k} \sqrt{k} \cdot g_j(n_j) \to 0$ as $N \to \infty$.

Central Limit Theorem

- Question: what happens when $k, N \to \infty$?
- $X_1, \ldots, X_N$ - independent from $P_1, \ldots, P_N$, θ_* – shared parameter (e.g., mean).
- $\bar{\theta}_1, \ldots, \bar{\theta}_k$ are asymptotically normal, $g_j(n_j)$ control approximation rate.
- $\max_{j=1,\ldots,k} \sqrt{k} \cdot g_j(n_j) \to 0$ as $N \to \infty$.
- $H_k := \left(\frac{1}{k} \sum_{j=1}^{k} \frac{1}{\sigma^{(j)}_{n_j}} \right)^{-1}$, and $\max\limits_{j \le k} \frac{H_k}{\sigma^{(j)}_{n_j} \sqrt{k}} \xrightarrow{N \to \infty} 0$

  [smallest of the variances in not too small]

Central Limit Theorem

- Question: what happens when $k, N \to \infty$?
- $X_1, \ldots, X_N$ - independent from $P_1, \ldots, P_N$, θ_* – shared parameter (e.g., mean).
- $\bar{\theta}_1, \ldots, \bar{\theta}_k$ are asymptotically normal, $g_j(n_j)$ control approximation rate.
- $\max_{j=1,\ldots,k} \sqrt{k} \cdot g_j(n_j) \to 0$ as $N \to \infty$.
- $H_k := \left(\frac{1}{k} \sum_{j=1}^{k} \frac{1}{\sigma_{n_j}^{(j)}} \right)^{-1}$, and $\max\limits_{j \le k} \frac{H_k}{\sigma_{n_j}^{(j)} \sqrt{k}} \xrightarrow{N \to \infty} 0$

Theorem

M. Under these assumptions,

$$\sqrt{k} \frac{\widehat{\theta}^{(k)} - \theta_*}{H_k} \xrightarrow{d} N\left(0, \frac{\pi}{2}\right).$$

Central Limit Theorem

- Question: what happens when $k, N \to \infty$?
- $X_1, \ldots, X_N$ - independent from $P_1, \ldots, P_N$, θ_* – shared parameter (e.g., mean).
- $\bar{\theta}_1, \ldots, \bar{\theta}_k$ are asymptotically normal, $g_j(n_j)$ control approximation rate.
- $\max_{j=1,\ldots,k} \sqrt{k} \cdot g_j(n_j) \to 0$ as $N \to \infty$.
- $H_k := \left(\frac{1}{k} \sum_{j=1}^{k} \frac{1}{\sigma_{n_j}^{(j)}} \right)^{-1}$, and $\max\limits_{j \le k} \frac{H_k}{\sigma_{n_j}^{(j)} \sqrt{k}} \xrightarrow{N \to \infty} 0$

Theorem

M. Under these assumptions,

$$\sqrt{k} \frac{\widehat{\theta}^{(k)} - \theta_*}{H_k} \xrightarrow{d} N\left(0, \frac{\pi}{2}\right).$$

In the i.i.d. scenario + equal group size, $\sigma_{n_j}^{(j)} = \frac{\sigma}{\sqrt{n}}$, *hence*

$$\sqrt{n}\left(\widehat{\theta}^{(k)} - \theta_*\right) \xrightarrow{d} N\left(0, \frac{\pi}{2}\sigma^2\right).$$

Multivariate case

- Assume that $\theta_* \in \mathbb{R}^d$.

Multivariate case

- Assume that $\theta_* \in \mathbb{R}^d$.
- $\widehat{\theta}^{(k)} := \text{argmin}_{z \in \mathbb{R}^m} \sum_{j=1}^{k} \left\| z - \bar{\theta}_j \right\|_{\mathbf{1}}$.

Multivariate case

- Assume that $\theta_* \in \mathbb{R}^d$.
- $\widehat{\theta}^{(k)} := \text{argmin}_{z \in \mathbb{R}^m} \sum_{j=1}^{k} \left\| z - \bar{\theta}_j \right\|_{\mathbf{1}}$.
- Only asymptotic normality of each coordinate is assumed.

Multivariate case

- Assume that $\theta_* \in \mathbb{R}^d$.
- $\widehat{\theta}^{(k)} := \operatorname{argmin}_{z \in \mathbb{R}^m} \sum_{j=1}^{k} \left\| z - \bar{\theta}_j \right\|_1$.
- Only asymptotic normality of each coordinate is assumed.
- For the mean estimation problem, we obtain that for all $i \leq d$,

$$\left| \widehat{\theta}_i^{(k)} - \theta_{*,i} \right| \leq C \cdot \sigma_i \left(\frac{\mathbb{E}|X_{1,i} - \theta_{*,i}|^{2+\delta} / \sigma_i^{2+\delta}}{n^{\frac{1+\delta}{2}}} + \sqrt{\frac{s}{N}} \right).$$

with probability at least $1 - 2de^{-4s}$.

Multivariate case

- Assume that $\theta_* \in \mathbb{R}^d$.
- $\widehat{\theta}^{(k)} := \operatorname{argmin}_{z \in \mathbb{R}^m} \sum_{j=1}^{k} \left\| z - \bar{\theta}_j \right\|_1$.
- Only asymptotic normality of each coordinate is assumed.
- For the mean estimation problem, we obtain that for all $i \leq d$,

$$\left| \widehat{\theta}_i^{(k)} - \theta_{*,i} \right| \leq C \cdot \sigma_i \left(\frac{\mathbb{E}|X_{1,i} - \theta_{*,i}|^{2+\delta} / \sigma_i^{2+\delta}}{n^{\frac{1+\delta}{2}}} + \sqrt{\frac{s}{N}} \right).$$

  with probability at least $1 - 2de^{-4s}$.
- Implies that

$$\left\| \widehat{\theta}_i^{(k)} - \theta_{*,i} \right\|_2 \leq C\sqrt{\operatorname{tr} \Sigma} \left(\max_i \frac{\mathbb{E}|X_{1,i} - \theta_{*,i}|^{2+\delta} / \sigma_i^{2+\delta}}{n^{\frac{1+\delta}{2}}} + \sqrt{\frac{s}{N}} \right).$$

Some questions

- Extensions to other versions of the multivariate median?
  (with optimal dependence on the dimension)
  E.g., spectral norm for matrices.

Some questions

- Extensions to other versions of the multivariate median?
  (with optimal dependence on the dimension)
  E.g., spectral norm for matrices.
- Extensions to empirical risk minimization: replace $\operatorname{argmin}_{f \in \mathcal{F}} P_n(f)$ by

$$\operatorname*{argmin}_{f \in \mathcal{F}} \mathrm{MOM}_k(f)$$

  For instance, $\mathcal{F} = \{\|z - \cdot\|_2^2\}$ gives mean estimation.

  [Existing work by Joly, Lecue, Lerasle, Lugosi, Mendelson]

Covariance estimation

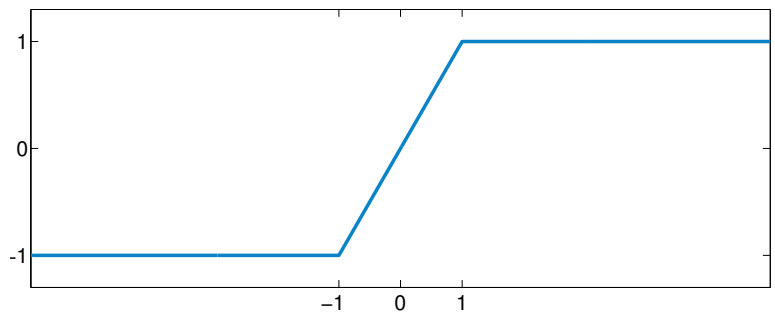
Huber's estimator

$X_1, \ldots, X_N$ - i.i.d., mean θ_*, variance σ^2.

Huber's estimator

$X_1, \ldots, X_N$ - i.i.d., mean θ_*, variance σ^2. Set

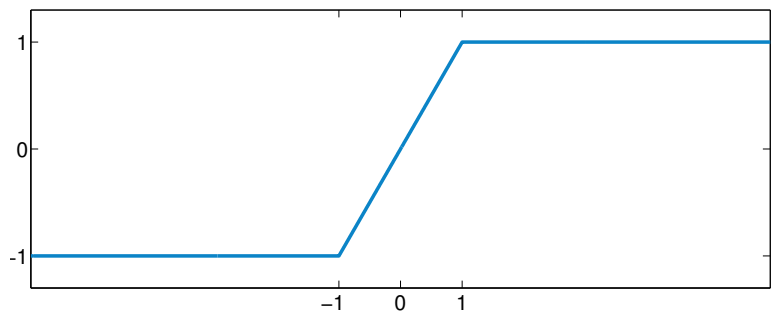
$$\psi(x) = (|x| \wedge 1)\text{sign}(x)$$



Huber's estimator

$X_1, \ldots, X_N$ - i.i.d., mean θ_*, variance σ^2. Set

$$\psi(x) = (|x| \wedge 1)\text{sign}(x)$$



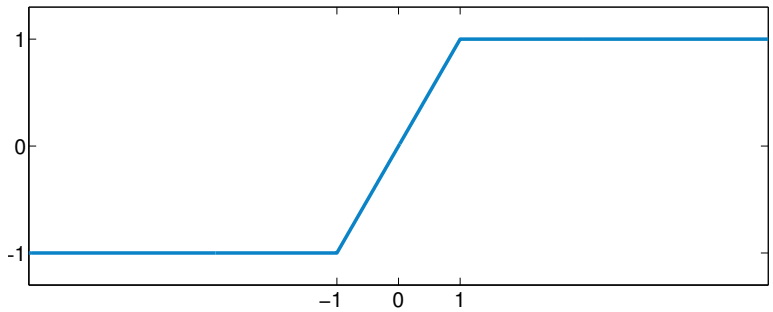
Let $\tau > 0$, and define $\hat{\theta} := \hat{\theta}(\tau)$ via

$$\sum_{j=1}^{N} \psi\left(\tau(X_j - \hat{\theta})\right) = 0.$$

Huber's estimator

$X_1, \ldots, X_N$ - i.i.d., mean θ_*, variance σ^2. Set

$$\psi(x) = (|x| \wedge 1)\text{sign}(x)$$



Let $\tau > 0$, and define $\hat{\theta} := \hat{\theta}(\tau)$ via

$$\sum_{j=1}^{N} \psi\left(\tau(X_j - \hat{\theta})\right) = 0.$$

Equivalent to minimizing Huber's loss

$$\hat{\theta} = \underset{\theta \in \mathbb{R}}{\text{argmin}} \sum_{j=1}^{N} H\left(\tau(X_j - \theta)\right)$$

Huber's estimator

Theoretical guarantees (O. Catoni '12): set $\tau = \sqrt{\frac{2\log(1/\alpha)}{N}}\frac{1}{\sigma}$. Then

$$\left|\hat{\theta}(\tau) - \theta_*\right| \leq \left(\sqrt{2} + o_N(1)\right)\sigma\sqrt{\frac{\log(1/\alpha)}{N}}$$

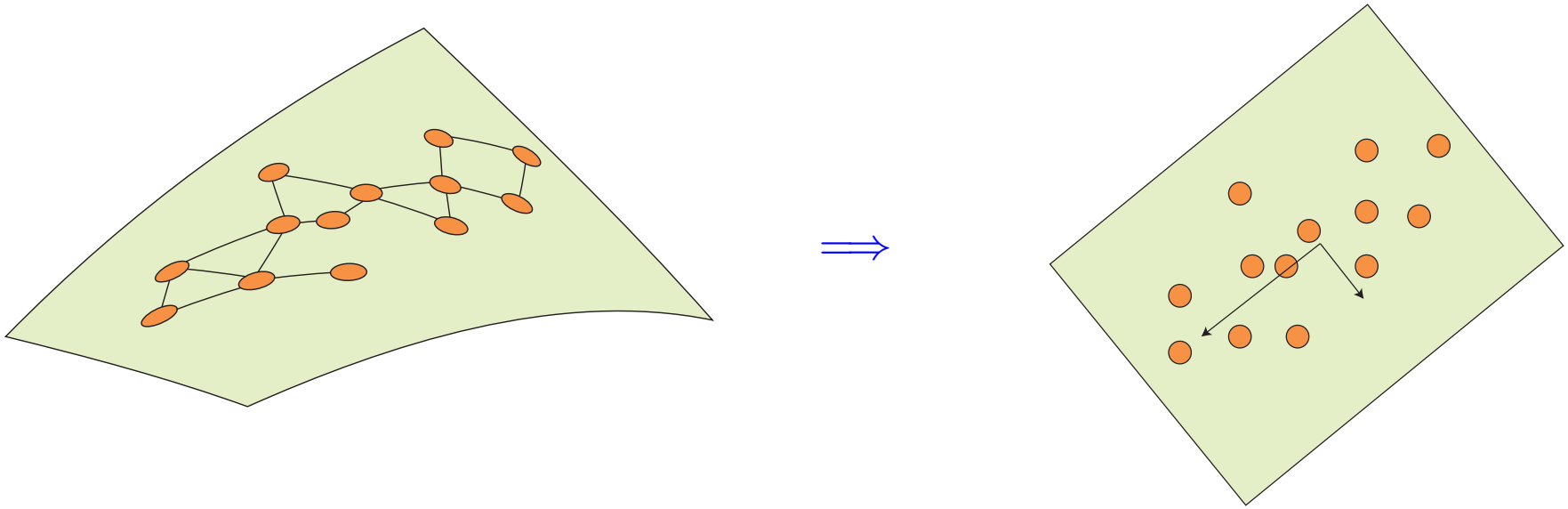
with probability $\geq 1 - 2\alpha$.

Extensions to higher dimensions

- Extend the method beyond the univariate case?

Extensions to higher dimensions

- Extend the method beyond the univariate case?
- Motivation: Covariance estimation and Principal Component Analysis (PCA)



Extensions to higher dimensions

- Motivation: Covariance estimation and Principal Component Analysis (PCA)
- Mathematical framework:

$$Y_1, \ldots, Y_N \in \mathbb{R}^d, \text{ i.i.d. } \mathbb{E}Y_j = \mu, \mathbb{E}(Y_j - \mu)(Y_j - \mu)^T = \Sigma,$$

$\mathbb{E}\,\|Y_j\|_2^4 < \infty$. No additional assumptions.

Extensions to higher dimensions

- Motivation: Covariance estimation and Principal Component Analysis (PCA)
- Mathematical framework:

$$Y_1, \ldots, Y_N \in \mathbb{R}^d, \text{ i.i.d. } \mathbb{E}Y_j = \mu, \mathbb{E}(Y_j - \mu)(Y_j - \mu)^T = \Sigma,$$
$$\mathbb{E}\,\|Y_j\|_2^4 < \infty. \text{ No additional assumptions.}$$

- Goal: construct $\hat{\Sigma}$, an estimator of Σ, such that

$$\underbrace{\left\|\hat{\Sigma} - \Sigma\right\|}_{\text{spectral norm}}$$

is small with high probability.

Extensions to higher dimensions

- Motivation: Covariance estimation and Principal Component Analysis (PCA)
- Mathematical framework:

$$Y_1, \dots, Y_N \in \mathbb{R}^d, \text{ i.i.d. } \mathbb{E}Y_j = \mu, \mathbb{E}(Y_j - \mu)(Y_j - \mu)^T = \Sigma,$$
$$\mathbb{E}\,\|Y_j\|_2^4 < \infty. \text{ No additional assumptions.}$$

- Goal: construct $\hat{\Sigma}$, an estimator of Σ, such that

$$\underbrace{\left\|\hat{\Sigma} - \Sigma\right\|}_{\text{spectral norm}}$$

  is small with high probability.
- The sample covariance matrix

$$\tilde{\Sigma}_N = \frac{1}{n-1}\sum_{j=1}^{N}(Y_j - \bar{Y}_N)(Y_j - \bar{Y}_N)^T$$

  is sensitive to outliers/heavy tails.

Extensions to higher dimensions

- Naive approach: apply Huber's estimator coordinatewise. Makes the bound
  - dimension-dependent
  - not invariant with respect to a change of coordinates.

Extensions to higher dimensions

- Naive approach: apply Huber's estimator coordinatewise.
  Makes the bound
  - dimension-dependent
  - not invariant with respect to a change of coordinates.
- Alternatives: Tukey's depth, Tyler's M-estimator, Maronna's M-estimator, Kendall's tau:
  - Guarantees are limited to special classes of distributions (e.g., elliptically symmetric).

Matrix functions

$f : \mathbb{R} \mapsto \mathbb{R}$, $A = A^T = U \Lambda U^T$, then

$$f(A) = U f(\Lambda) U^T, \quad f(\Lambda) = f\left(\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix}\right) = \begin{pmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{pmatrix}$$

Construction of the estimator

- $Y \in \mathbb{R}^d$, $Y_1, \ldots, Y_N \in \mathbb{R}^d$ – i.i.d. copies of Y, μ is the mean, Σ is the covariance matrix,

$$\mathbb{E}\|Y\|^4 < \infty, \qquad \textit{No additional assumptions}.$$

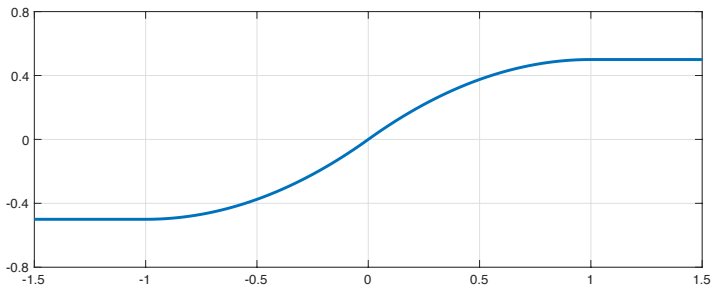
Construction of the estimator

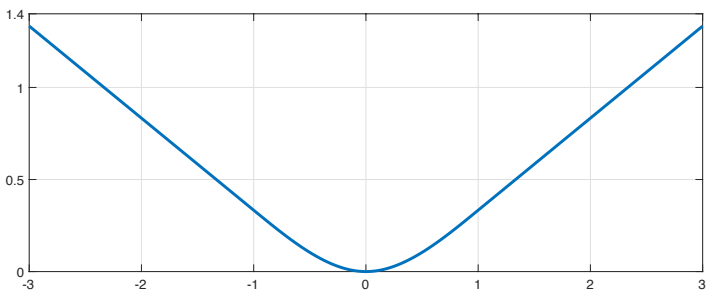
- $Y \in \mathbb{R}^d$, $Y_1, \dots, Y_N \in \mathbb{R}^d$ – i.i.d. copies of Y, μ is the mean, Σ is the covariance matrix,

$$\mathbb{E}\|Y\|^4 < \infty, \qquad \textit{No additional assumptions}.$$

- Set $\Psi'(x) = \psi(x) = \begin{cases} 1/2, & x > 1, \\ x - x^2/2, & x \in [0,1], \\ x + x^2/2, & x \in [-1,0), \\ -1/2, & x < -1. \end{cases}$

$$\Big[\text{like Huber's loss + operator Lipschitz}\Big]$$





Construction of the estimator

- $Y \in \mathbb{R}^d$, $Y_1, \ldots, Y_N \in \mathbb{R}^d$ – i.i.d. copies of Y, μ is the mean, Σ is the covariance matrix,

$$\mathbb{E}\|Y\|^4 < \infty, \qquad \textit{No additional assumptions}.$$

- Set $\Psi'(x) = \psi(x) = \begin{cases} 1/2, & x > 1, \\ x - x^2/2, & x \in [0,1], \\ x + x^2/2, & x \in [-1,0), \\ -1/2, & x < -1. \end{cases}$

$$\Big[\text{like Huber's loss + operator Lipschitz}\Big]$$

- Observe that

$$\Sigma = \frac{1}{2}\mathbb{E}\left[(Y_1 - Y_2)(Y_1 - Y_2)^T\right]$$

Construction of the estimator

- $Y \in \mathbb{R}^d$, $Y_1, \ldots, Y_N \in \mathbb{R}^d$ – i.i.d. copies of Y, μ is the mean, Σ is the covariance matrix,

$$\mathbb{E}\|Y\|^4 < \infty, \qquad \textit{No additional assumptions}.$$

- Set $\Psi'(x) = \psi(x) = \begin{cases} 1/2, & x > 1, \\ x - x^2/2, & x \in [0,1], \\ x + x^2/2, & x \in [-1,0), \\ -1/2, & x < -1. \end{cases}$

$$\Big[\text{like Huber's loss + operator Lipschitz}\Big]$$

- Observe that

$$\Sigma = \frac{1}{2}\mathbb{E}\left[(Y_1 - Y_2)(Y_1 - Y_2)^T\right]$$

- The sample covariance is then

$$\widetilde{\Sigma} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}$$

$$= \frac{1}{N-1} \sum_{j=1}^{N} (Y_j - \bar{Y}_N)(Y_j - \bar{Y}_N)^T$$

Construction of the estimator

- The sample covariance is

$$\widetilde{\Sigma} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}$$

This an example of a U-statistic.

Construction of the estimator

- The sample covariance is

$$\widetilde{\Sigma} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}$$

  This an example of a U-statistic.

- Equivalently,

$$\underbrace{\frac{1}{N(N-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}}_{\text{average}} = \operatorname*{argmin}_{S \in \mathbb{R}^{d \times d}} \sum_{i \neq j} \left\| \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right\|_{\mathrm{F}}^2$$

Construction of the estimator

- The sample covariance is

$$\widetilde{\Sigma} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}$$

  This an example of a U-statistic.

- Equivalently,

$$\underbrace{\frac{1}{N(N-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}}_{\text{average}} = \operatorname*{argmin}_{S \in \mathbb{R}^{d \times d}} \sum_{i \neq j} \left\| \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right\|_{\mathrm{F}}^2$$

$$= \operatorname*{argmin}_{S \in \mathbb{R}^{d \times d}} \operatorname{Trace} \left[\sum_{i \neq j} \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right)^2 \right]$$

Construction of the estimator

- The sample covariance is

$$\widetilde{\Sigma} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}$$

  This an example of a U-statistic.

- Equivalently,

$$\underbrace{\frac{1}{N(N-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}}_{\text{average}} = \operatorname*{argmin}_{S \in \mathbb{R}^{d \times d}} \sum_{i \neq j} \left\| \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right\|_{\mathrm{F}}^2$$

$$= \operatorname*{argmin}_{S \in \mathbb{R}^{d \times d}} \operatorname{Trace} \left[\sum_{i \neq j} \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right)^2 \right]$$

- Replace quadratic loss by (rescaled) loss $\Psi(x)$: let $\theta > 0$ [small constant] , and define

$$\widehat{\Sigma} = \operatorname*{argmin}_{S \in \mathbb{R}^{d \times d}} \left[\operatorname{Trace} \sum_{i \neq j} \Psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

$$\widehat{\Sigma} = \underset{S \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \left[\operatorname{Trace} \sum_{i \neq j} \Psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

Equivalent to

$$\frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{\theta} \psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - \widehat{\Sigma} \right) \right) = 0_{d \times d}.$$

Remark

Since Ψ is convex, $\widehat{\Sigma}$ can be obtained via the gradient descent iteration

$$\widehat{\Sigma}_k = \widehat{\Sigma}_{k-1} + \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{\theta} \psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - \widehat{\Sigma}_{k-1} \right) \right)$$

Theoretical guarantees

- The effective rank of a nonnegative definite matrix A is

$$\mathrm{R}(A) = \frac{\operatorname{tr} A}{\|A\|}$$

Theoretical guarantees

- The effective rank of a nonnegative definite matrix A is

$$\mathrm{R}(A) = \frac{\operatorname{tr} A}{\|A\|}$$

- $H(Y_1, Y_2) = \frac{(Y_1 - Y_2)(Y_1 - Y_2)^T}{2}$

Theoretical guarantees

- The effective rank of a nonnegative definite matrix A is

$$\mathrm{R}(A) = \frac{\operatorname{tr} A}{\|A\|}$$

- $H(Y_1, Y_2) = \frac{(Y_1 - Y_2)(Y_1 - Y_2)^T}{2}$
- Set

$$\mathrm{R}_H := \mathrm{R}\left(\underbrace{\mathbb{E}\left(H(Y_1, Y_2) - \mathbb{E}H\right)^2}_{\text{“variance matrix”}} \right)$$

Theoretical guarantees

- The effective rank of a nonnegative definite matrix A is

$$\mathrm{R}(A) = \frac{\operatorname{tr} A}{\|A\|}$$

- $H(Y_1, Y_2) = \frac{(Y_1 - Y_2)(Y_1 - Y_2)^T}{2}$
- Set

$$\mathrm{R}_H := \mathrm{R}\left(\underbrace{\mathbb{E}\left(H(Y_1, Y_2) - \mathbb{E}H\right)^2}_{\text{“variance matrix”}} \right)$$

Theorem

Fix $\alpha > 0$. Assume that $\sigma_0^2 \geq \|\mathbb{E}\left(H(Y_1, Y_2) - \mathbb{E}H\right)^2\|$, and let $\theta = \sqrt{\frac{4\log(4d/\alpha)}{N}}\frac{1}{\sigma_0}$. If $\frac{R_H \log(d/\alpha)}{N} \leq \frac{1}{200}$, then

$$\left\|\widehat{\Sigma} - \Sigma\right\| \leq 23\sqrt{2}\sigma_0\sqrt{\frac{\log(4d/\alpha)}{N}}$$

with probability $\geq 1 - \alpha$.

Theoretical guarantees

Theorem

Fix $\alpha > 0$. *Assume that* $\sigma_0^2 \geq \| \mathbb{E} \left(H(Y_1, Y_2) - \mathbb{E}H \right)^2 \|$, *and let* $\theta = \sqrt{\frac{4 \log(4d/\alpha)}{N}} \frac{1}{\sigma_0}$. *If* $\frac{R_H \log(d/\alpha)}{N} \leq \frac{1}{200}$, *then*

$$\left\| \widehat{\Sigma} - \Sigma \right\| \leq 23\sqrt{2}\sigma_0 \sqrt{\frac{\log(4d/\alpha)}{N}}$$

with probability $\geq 1 - \alpha$.

- Proof of the bound is based on the analysis of the gradient descent scheme

$$\widehat{\Sigma}_k = \widehat{\Sigma}_{k-1} + \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{\theta} \psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - \widehat{\Sigma}_{k-1} \right) \right)$$

Implications for the covariance estimation problem

$$\widehat{\Sigma} = \underset{S \in \mathbb{R}^{d \times d}}{\text{argmin}} \left[\text{Trace} \sum_{i \neq j} \Psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

Implications for the covariance estimation problem

$$\widehat{\Sigma} = \underset{S \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \left[\operatorname{Trace} \sum_{i \neq j} \Psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

- Let

$$K^2 := \sup_{v : \|v\|_2 = 1} \frac{\mathbb{E}\langle Y - \mathbb{E}Y, v \rangle^4}{\left[\mathbb{E}\langle Y - \mathbb{E}Y, v \rangle^2 \right]^2}$$

Implications for the covariance estimation problem

$$\widehat{\Sigma} = \operatorname*{argmin}_{S \in \mathbb{R}^{d \times d}} \left[\operatorname{Trace} \sum_{i \neq j} \Psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

- Let

$$K^2 := \sup_{v : \|v\|_2 = 1} \frac{\mathbb{E}\langle Y - \mathbb{E}Y, v \rangle^4}{\left[\mathbb{E}\langle Y - \mathbb{E}Y, v \rangle^2 \right]^2}$$

Corollary

Fix $\alpha > 0$. *Assume that* $\sigma_0^2 = K\, R(\Sigma) \|\Sigma\|^2$, *and let* $\theta = \sqrt{\frac{4 \log(4d/\alpha)}{N}} \frac{1}{\sigma_0}$. *If* $\frac{r(\Sigma) \log(4d/\alpha)}{N} \leq \frac{1}{200}$, *then*

$$\left\| \widehat{\Sigma} - \Sigma \right\| \leq 23 \sqrt{2K}\, \|\Sigma\| \sqrt{R(\Sigma) \frac{\log(4d/\alpha)}{N}}$$

with probability $\geq 1 - \alpha$.