Uniform bounds for robust mean estimators

Stas Minsker

Department of Mathematics, USC

February 14

ITA Workshop, Robust Learning

Noisy and corrupted data

is one of the challenges in contemporary statistics and data science:

Noisy and corrupted data
is one of the challenges in contemporary statistics and data science:

- Presence of outliers of unknown nature:

  $\Longrightarrow$ requires algorithms that are robust.

Noisy and corrupted data
is one of the challenges in contemporary statistics and data science:

- Presence of outliers of unknown nature:

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

Noisy and corrupted data
is one of the challenges in contemporary statistics and data science:

- 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 "noisy" data is via heavy-tailed distributions.

Noisy and corrupted data
is one of the challenges in contemporary statistics and data science:

- 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 “noisy” data is via heavy-tailed distributions.
- For the purpose of this talk, a random variable X has heavy-tailed distribution if

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

  for some $r > 0$ (for example, $r = 2.1$).

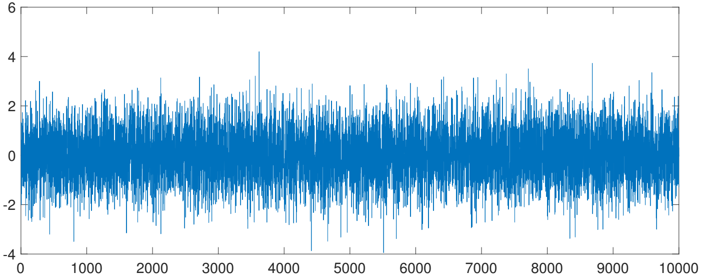


Figure: Standard normal distribution.

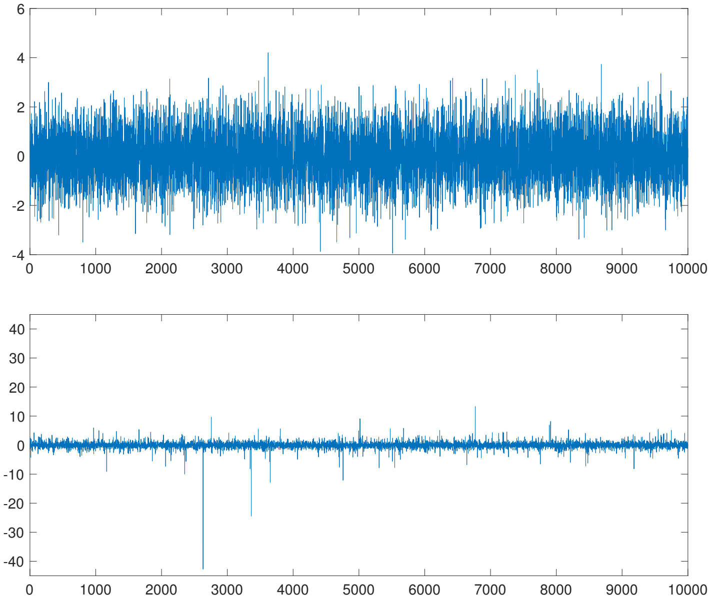


Figure: Student's t-distribution with 3 d.f.

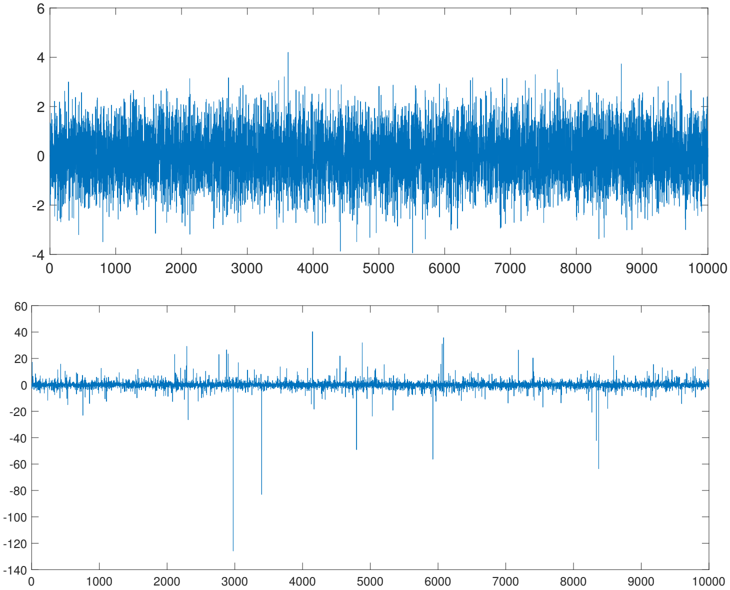


Figure: Student's t-distribution with 2.1 d.f.

Question: how to estimate the mean?

- Motivation: risk minimization of the form

  approximate the minimizer of the "risk" $\mathbb{E}\ell(Y, f(X))$ over the class $\mathcal{F}$.

Question: how to estimate the mean?

- Motivation: risk minimization of the form

  approximate the minimizer of the "risk" $\mathbb{E}\ell(Y, f(X))$ over the class $\mathcal{F}$.

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

Question: how to estimate the mean?

- Motivation: risk minimization of the form

  approximate the minimizer of the "risk" $\mathbb{E}\ell(Y, f(X))$ over the class $\mathcal{F}$.

- Benchmark: 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

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

  similar inequality holds for sub-Gaussian distributions.

Question: how to estimate the mean?

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

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

on Π – possibly asymmetric, with heavy tails.

Question: how to estimate the mean?

- 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:

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

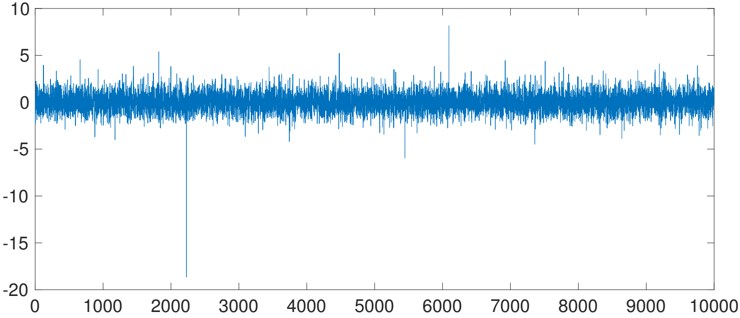


Figure: Rescaled Sample Means of Student's t-distribution with **3 d.f.**

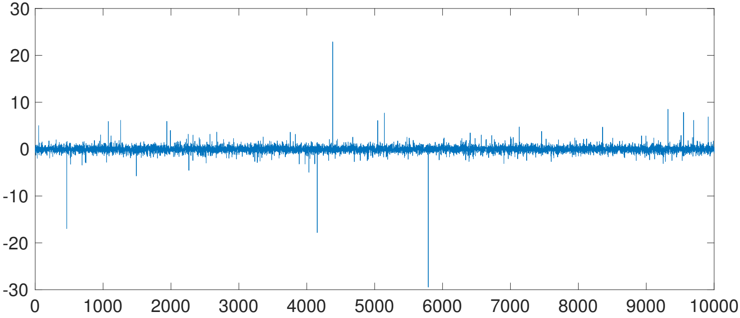


Figure: Rescaled Sample Means of Student's t-distribution with **2.1 d.f.**

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:

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

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:

$$\overbrace{X_1, \ldots, X_{|G_1|}}^{G_1} \;\cdots\cdots\; \overbrace{X_{N-|G_k|+1}, \ldots, X_N}^{G_k}$$

$$\bar{\mu}_1 := \frac{1}{|G_1|} \sum_{X_i \in G_1} X_i \qquad \bar{\mu}_k := \frac{1}{|G_k|} \sum_{X_i \in G_k} X_i$$

$$\widehat{\mu}^{(k)} := \mathrm{median}(\bar{\mu}_1, \ldots, \bar{\mu}_k)$$

- Claim:

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

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:
- Claim:

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

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:
- Claim:

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

- Has recently been extended to multivariate mean and covariance estimation, empirical risk minimization, U-statistics.
  Quickly growing body of work: G. Chinot, L. Devroye, E. Joly, G. Lecué, M. Lerasle, G. Lugosi, T. Matthieu, S. Mendelson, R. Oliveira, S. Hopkins, N. Zhivotovsky.

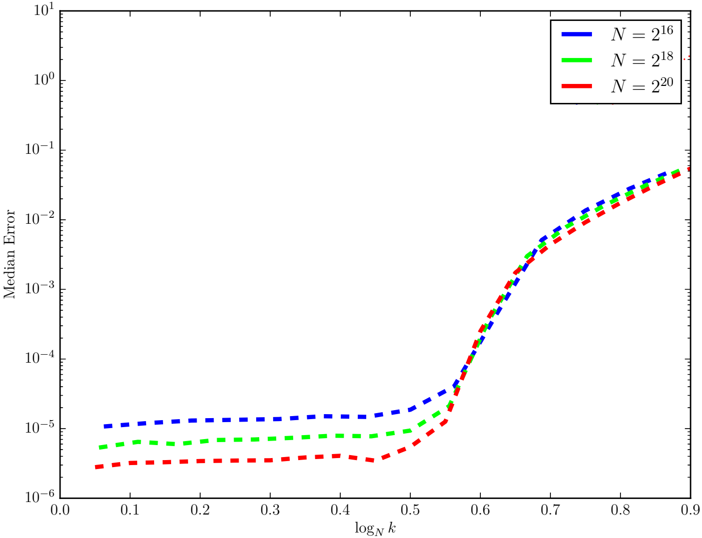
Question: how to estimate the mean?

- 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:
- Claim:

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

- Has recently been extended to multivariate mean and covariance estimation, empirical risk minimization, U-statistics.
  Quickly growing body of work: G. Chinot, L. Devroye, E. Joly, G. Lecué, M. Lerasle, G. Lugosi, T. Matthieu, S. Mendelson, R. Oliveira, S. Hopkins, N. Zhivotovsky.
- Similar results were obtained by J. Fan, W.-X. Zhou, Z. Ren, O. Catoni, I. Giulini using different estimation techniques.

Perfomance as k changes



Result so far:

$$\Pr\left(|\widehat{\mu}^{(k)} - \mu| \geq C\,\sigma\sqrt{\frac{k}{N}}\right) \leq e^{-k} := \alpha$$

- Need to recalculate the estimator for different values of confidence parameter α. Can one "decouple" k and α?

Result so far:

$$\Pr\left(|\widehat{\mu}^{(k)} - \mu| \geq C\,\sigma\sqrt{\frac{k}{N}}\right) \leq e^{-k} := \alpha$$

- Need to recalculate the estimator for different values of confidence parameter α. Can one "decouple" k and α?
- How to choose k? Typically, want k as large as possible.

Result so far:

$$\Pr\left(|\widehat{\mu}^{(k)} - \mu| \geq C\,\sigma\sqrt{\frac{k}{N}}\right) \leq e^{-k} := \alpha$$

- Need to recalculate the estimator for different values of confidence parameter α. Can one "decouple" k and α?
- How to choose k? Typically, want k as large as possible.
- Limiting distribution: what happens when $k, N \to \infty$?

Result so far:

$$\Pr\left(|\widehat{\mu}^{(k)} - \mu| \geq C\,\sigma\sqrt{\frac{k}{N}}\right) \leq e^{-k} := \alpha$$

- Need to recalculate the estimator for different values of confidence parameter α. Can one "decouple" k and α?
- How to choose k? Typically, want k as large as possible.
- Limiting distribution: what happens when $k, N \to \infty$?
- Assume that many means need to be estimated simultaneously. Uniform deviation bounds?

Result so far:

$$\Pr\left(|\widehat{\mu}^{(k)} - \mu| \geq C\,\sigma\sqrt{\frac{k}{N}}\right) \leq e^{-k} := \alpha$$

- Need to recalculate the estimator for different values of confidence parameter α. Can one "decouple" k and α?
- How to choose k? Typically, want k as large as possible.
- Limiting distribution: what happens when $k, N \to \infty$?
- Assume that many means need to be estimated simultaneously. Uniform deviation bounds?
- Robust estimator that does not depend on the random partition of the index set?

Result so far:

$$\Pr\left(|\widehat{\mu}^{(k)} - \mu| \geq C\,\sigma\sqrt{\frac{k}{N}}\right) \leq e^{-k} := \alpha$$

- Need to recalculate the estimator for different values of confidence parameter α. Can one "decouple" k and α?
- How to choose k? Typically, want k as large as possible.
- Limiting distribution: what happens when $k, N \to \infty$?
- Assume that many means need to be estimated simultaneously. Uniform deviation bounds?
- Robust estimator that does not depend on the random partition of the index set?
- Algorithms for robust Empirical Risk Minimization?

Connections between symmetry and robustness

1. If the distribution P is symmetric, then its center of symmetry $\theta(P)$ can be approximated by a robust estimator with a high breakdown point, e.g. a robust M-estimator

$$\widehat{\theta} := \underset{z \in \mathbb{R}}{\operatorname{argmin}} \sum_{j=1}^{N} \rho\left(z - X_j\right).$$

Connections between symmetry and robustness

1. If the distribution P is symmetric, then its center of symmetry $\theta(P)$ can be approximated by a robust estimator with a high breakdown point, e.g. a robust M-estimator
$$\widehat{\theta} := \operatorname*{argmin}_{z \in \mathbb{R}} \sum_{j=1}^{N} \rho\left(z - X_j\right).$$
2. In order to obtain a robust estimator of a parameter $\theta(P)$ of (not necessarily symmetric) distribution P based on the i.i.d. sample $X_1, \ldots, X_N$, create a new sample such that
   - (i) it is governed by an approximately symmetric distribution;
   - (ii) the center of symmetry of this distribution is close to $\theta(P)$.

Connections between symmetry and robustness

1. If the distribution P is symmetric, then its center of symmetry $\theta(P)$ can be approximated by a robust estimator with a high breakdown point, e.g. a robust M-estimator

$$\widehat{\theta} := \underset{z\in\mathbb{R}}{\operatorname{argmin}} \sum_{j=1}^{N} \rho\left(z - X_j\right).$$

2. In order to obtain a robust estimator of a parameter $\theta(P)$ of (not necessarily symmetric) distribution P based on the i.i.d. sample $X_1, \ldots, X_N$, create a new sample such that
   - (i) it is governed by an approximately symmetric distribution;
   - (ii) the center of symmetry of this distribution is close to $\theta(P)$.

How does one create such a "new sample"? A possible approach is based on the fact that

as sample size grows, the summary statistics of the data become asymptotically normal, hence asymptotically symmetric. Examples: sample mean, MLE.

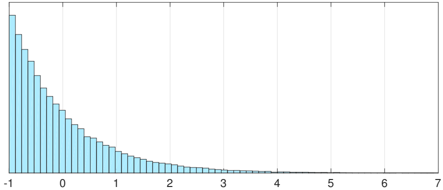


Figure: Centered exponential distribution

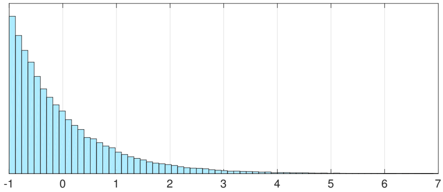


Figure: Centered exponential distribution

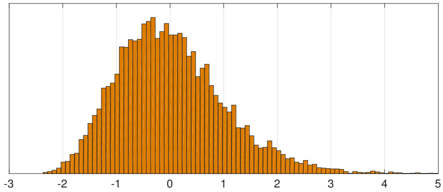


Figure: Rescaled sample means with $n = 10$.

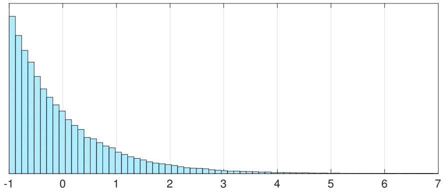


Figure: Centered exponential distribution

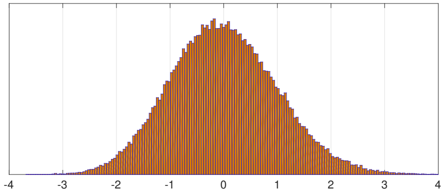


Figure: Rescaled sample means with $n = 100$.

Robust estimators of the mean

- Split the sample into k groups $G_1, \ldots, G_k$ of size $n_j = |G_j|$ each:

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

Robust estimators of the mean

- Split the sample into k groups $G_1, \ldots, G_k$ of size $n_j = |G_j|$ each:

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

- ρ is convex, even function such that $\rho(z) \to \infty$ as $|z| \to \infty$ and $\|\rho'\|_\infty < \infty$.

Robust estimators of the mean

- Split the sample into k groups $G_1, \dots, G_k$ of size $n_j = |G_j|$ each:

$$\underbrace{\overbrace{X_1, \dots, X_{|G_1|}}^{G_1}}_{\bar{\mu}_1 := \frac{1}{|G_1|} \sum_{X_i \in G_1} X_i} \;\cdots\cdots\; \underbrace{\overbrace{X_{N-|G_k|+1}, \dots, X_N}^{G_k}}_{\bar{\mu}_k := \frac{1}{|G_k|} \sum_{X_i \in G_k} X_i}$$

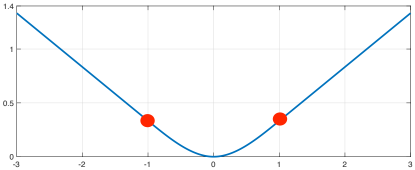
- ρ is convex, even function such that $\rho(z) \to \infty$ as $|z| \to \infty$ and $\|\rho'\|_\infty < \infty$.
- $\widehat{\mu}^{(k)} := \text{argmin}_{z \in \mathbb{R}} \sum_{j=1}^{k} \rho\left(\sqrt{n_j}\, \frac{\bar{\mu}_j - z}{\Delta}\right)$, where $\Delta > 0$.

Robust estimators of the mean

- Split the sample into k groups $G_1, \ldots, G_k$ of size $n_j = |G_j|$ each:

$$\underbrace{\overbrace{X_1, \ldots, X_{|G_1|}}^{G_1}}_{\bar{\mu}_1 := \frac{1}{|G_1|}\sum\limits_{X_i \in G_1} X_i} \; \cdots\cdots \; \underbrace{\overbrace{X_{N-|G_k|+1}, \ldots, X_N}^{G_k}}_{\bar{\mu}_k := \frac{1}{|G_k|}\sum\limits_{X_i \in G_k} X_i}$$

- ρ is convex, even function such that $\rho(z) \to \infty$ as $|z| \to \infty$ and $\|\rho'\|_\infty < \infty$.
- $\widehat{\mu}^{(k)} := \operatorname{argmin}_{z \in \mathbb{R}} \sum_{j=1}^{k} \rho\left(\sqrt{n_j}\, \frac{\bar{\mu}_j - z}{\Delta}\right)$, where $\Delta > 0$.
- Examples:
  1. $\rho(x) = |x|$ yields the median-of-means estimator.
  2. $\rho(x) =$ Huber's loss:



Non-asymptotic guarantees

- $X_1, \ldots, X_N$ are i.i.d., with mean μ and variance σ^2.
- Will assume that $n_1 = \ldots = n_k = n$ during the talk.
- $\Phi(t)$ - distribution function of $N(0,1)$, and

$$g(n) := \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\sqrt{n} \frac{\bar{X}_n - \mathbb{E}X}{\sqrt{\mathrm{Var}(X)}} \le t \right) - \Phi(t) \right|.$$

Non-asymptotic guarantees

- $X_1, \ldots, X_N$ are i.i.d., with mean μ and variance σ^2.
- Will assume that $n_1 = \ldots = n_k = n$ during the talk.
- $\Phi(t)$ - distribution function of $N(0,1)$, and

$$g(n) := \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\sqrt{n} \frac{\bar{X}_n - \mathbb{E}X}{\sqrt{\mathrm{Var}(X)}} \leq t \right) - \Phi(t) \right|.$$

- $\underline{c}, \bar{C} > 0$ are absolute constants,

$$\widetilde{\Delta} = \max(\Delta, \sigma)$$

Non-asymptotic guarantees

- $X_1, \ldots, X_N$ are i.i.d., with mean μ and variance σ^2.
- Will assume that $n_1 = \ldots = n_k = n$ during the talk.
- $\Phi(t)$ - distribution function of $N(0,1)$, and

$$g(n) := \sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\frac{\bar{X}_n - \mathbb{E}X}{\sqrt{\mathrm{Var}(X)}} \leq t \right) - \Phi(t) \right|.$$

- $\underline{c}, \bar{C} > 0$ are absolute constants,

$$\widetilde{\Delta} = \max(\Delta, \sigma)$$

Non-asymptotic guarantees

- $X_1, \ldots, X_N$ are i.i.d., with mean μ and variance σ^2.
- Will assume that $n_1 = \ldots = n_k = n$ during the talk.
- $\Phi(t)$ - distribution function of $N(0,1)$, and

$$g(n) := \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\sqrt{n} \frac{\bar{X}_n - \mathbb{E}X}{\sqrt{\mathrm{Var}(X)}} \leq t \right) - \Phi(t) \right|.$$

- $\underset{\bar{}}{c}, \bar{C} > 0$ are absolute constants,

$$\widetilde{\Delta} = \max(\Delta, \sigma)$$

Non-asymptotic guarantees

- $X_1, \ldots, X_N$ are i.i.d., with mean μ and variance σ^2.
- Will assume that $n_1 = \ldots = n_k = n$ during the talk.
- $\Phi(t)$ - distribution function of $N(0,1)$, and

$$g(n) := \sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n} \frac{\bar{X}_n - \mathbb{E}X}{\sqrt{\mathrm{Var}(X)}} \leq t \right) - \Phi(t) \right|.$$

- $\underline{c}, \bar{C} > 0$ are absolute constants,

$$\widetilde{\Delta} = \max(\Delta, \sigma)$$

Theorem (M., 2018)

For all $s > 0$ such that $\sqrt{\frac{s}{k}} + g(n) \leq \underline{c}(\rho)$,

$$\left| \widehat{\mu}^{(k)} - \mu \right| \leq \bar{C}(\rho)\, \widetilde{\Delta} \left(\sqrt{\frac{s}{N}} + g(n) \sqrt{\frac{k}{N}} \right)$$

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

Non-asymptotic guarantees

- $X_1, \ldots, X_N$ are i.i.d., with mean μ and variance σ^2.
- Will assume that $n_1 = \ldots = n_k = n$ during the talk.
- $\Phi(t)$ - distribution function of $N(0,1)$, and

$$g(n) := \sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\frac{\bar{X}_n - \mathbb{E}X}{\sqrt{\mathrm{Var}(X)}} \leq t \right) - \Phi(t) \right|.$$

- $\underline{c}, \bar{C} > 0$ are absolute constants,

$$\widetilde{\Delta} = \max(\Delta, \sigma)$$

Theorem (M., 2018)

For all $s > 0$ such that $\sqrt{\frac{s}{k}} + g(n) \leq \underline{c}(\rho)$,

$$\left| \widehat{\mu}^{(k)} - \mu \right| \leq \bar{C}(\rho)\, \widetilde{\Delta} \left(\sqrt{\frac{s}{N}} + \underbrace{g(n)\sqrt{\frac{k}{N}}}_{\text{"bias"}} \right)$$

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

Non-asymptotic guarantees

- $\Phi(t)$ - distribution function of $N(0,1)$, and
$$g(n) := \sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\frac{\bar{X}_n - \mathbb{E}X}{\sqrt{\text{Var}(X)}} \le t \right) - \Phi(t) \right|.$$
- $\underline{c}, \bar{C} > 0$ are absolute constants,
$$\widetilde{\Delta} = \max(\Delta, \sigma)$$

Theorem (M., 2018)

For all $s > 0$ such that $\sqrt{\frac{s}{k}} + g(n) \le \underline{c}(\rho)$,
$$\left| \widehat{\mu}^{(k)} - \mu \right| \le \bar{C}(\rho)\, \widetilde{\Delta} \left(\sqrt{\frac{s}{N}} + \underbrace{g(n)\sqrt{\frac{k}{N}}}_{\text{"bias"}} \right)$$
with probability at least $1 - 2e^{-s}$.

- Moreover, if $k \le C/g^2(n)$, then $\mathbb{E}\left| \widehat{\mu}^{(k)} - \mu \right| \le C(\rho) \frac{\widetilde{\Delta}}{\sqrt{N}}$.

Non-asymptotic guarantees

- $\Phi(t)$ - distribution function of $N(0,1)$, and
$$g(n) := \sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\frac{\bar{X}_n - \mathbb{E}X}{\sqrt{\mathrm{Var}(X)}} \le t \right) - \Phi(t) \right|.$$
- $\underline{c}, \bar{C} > 0$ are absolute constants,
$$\widetilde{\Delta} = \max(\Delta, \sigma)$$

Add $\mathcal{O}$ arbitrary (e.g., adversarially generated) outliers:

Theorem (M., 2018)

For all $\mathcal{O} \in \mathbb{N}$, $s > 0$ *such that* $\sqrt{\frac{s}{k}} + g(n) + \frac{\mathcal{O}}{k} \le \underline{c}(\rho)\left(1 - \frac{\mathcal{O}}{k}\right)$,

$$\left|\widehat{\mu}^{(k)} - \mu\right| \le \bar{C}(\rho)\,\widetilde{\Delta}\left(\sqrt{\frac{s}{N}} + g(n)\sqrt{\frac{k}{N}} + \frac{\mathcal{O}}{\sqrt{k}}\frac{1}{\sqrt{N}} \right)$$

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

Theorem (M., 2018)

For all $\mathcal{O} \in \mathbb{N}$, $s > 0$ *such that* $\|\rho'\|_\infty \left(\sqrt{\frac{s}{k}} + g(n) + \frac{\mathcal{O}}{k}\right) \leq \underline{c}\left(1 - \frac{\mathcal{O}}{k}\right)$,

$$\left|\widehat{\mu}^{(k)} - \mu\right| \leq \bar{C}\max(\Delta, \sigma)\,\|\rho'\|_\infty \left(\sqrt{\frac{s}{N}} + g(n)\sqrt{\frac{k}{N}} + \frac{\mathcal{O}}{\sqrt{k}}\frac{1}{\sqrt{N}}\right)$$

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

- For example, if $\mathbb{E}|X-\mu|^3 < \infty$, then $g(n) \lesssim \frac{\mathbb{E}|X-\theta_*|^3}{\sigma^3 n^{1/2}}$, and we get the bound

$$\left|\widehat{\mu}^{(k)} - \mu\right| \leq \bar{C}\,\Delta\,\|\rho'\|_\infty \left(\sqrt{\frac{s}{N}} + \frac{\mathbb{E}|X-\theta_*|^3}{\sigma^3}\frac{k}{N} + \frac{\mathcal{O}}{\sqrt{k}}\frac{1}{\sqrt{N}}\right)$$

that holds with probability $\geq 1 - 2e^{-s}$.

Theorem (M., 2018)

For all $\mathcal{O} \in \mathbb{N}$, $s > 0$ *such that* $\|\rho'\|_\infty \left(\sqrt{\frac{s}{k}} + g(n) + \frac{\mathcal{O}}{k}\right) \leq \underline{c}\left(1 - \frac{\mathcal{O}}{k}\right)$,

$$\left|\widehat{\mu}^{(k)} - \mu\right| \leq \bar{C} \max(\Delta, \sigma) \|\rho'\|_\infty \left(\sqrt{\frac{s}{N}} + g(n)\sqrt{\frac{k}{N}} + \frac{\mathcal{O}}{\sqrt{k}} \frac{1}{\sqrt{N}}\right)$$

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

- For example, if $\mathbb{E}|X - \mu|^3 < \infty$, then $g(n) \lesssim \frac{\mathbb{E}|X - \theta_*|^3}{\sigma^3 n^{1/2}}$, and we get the bound

$$\left|\widehat{\mu}^{(k)} - \mu\right| \leq \bar{C}\,\Delta \|\rho'\|_\infty \left(\sqrt{\frac{s}{N}} + \frac{\mathbb{E}|X - \theta_*|^3}{\sigma^3} \frac{k}{N} + \frac{\mathcal{O}}{\sqrt{k}} \frac{1}{\sqrt{N}}\right)$$

  that holds with probability $\geq 1 - 2e^{-s}$.
- If $\mathcal{O} = \varepsilon \cdot N$, then “optimal” $k \simeq \varepsilon^{2/3} N$ and resulting error is of order $\varepsilon^{2/3}$.

Asymptotic results

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

Asymptotic results

- Question: what happens when $k, N \to \infty$?
- Assume that $\sqrt{k} \cdot g(n) \to 0$ as $N \to \infty$ (if $\mathbb{E}|X - \mu|^{2+\delta} < \infty$, then $k = o\left(N^{\frac{\delta}{1+\delta}}\right)$ suffices).

Asymptotic results

- Question: what happens when $k, N \to \infty$?
- Assume that $\sqrt{k} \cdot g(n) \to 0$ as $N \to \infty$ (if $\mathbb{E}|X - \mu|^{2+\delta} < \infty$, then $k = o\left(N^{\frac{\delta}{1+\delta}}\right)$ suffices).
- $L(z) := \mathbb{E}\rho'\left(z + Z\right)$, where $Z \sim N(0, 1)$.

Asymptotic results

- Question: what happens when $k, N \to \infty$?
- Assume that $\sqrt{k} \cdot g(n) \to 0$ as $N \to \infty$ (if $\mathbb{E}|X - \mu|^{2+\delta} < \infty$, then $k = o\left(N^{\frac{\delta}{1+\delta}}\right)$ suffices).
- $L(z) := \mathbb{E}\rho'\left(z + Z\right)$, where $Z \sim N(0, 1)$.
- $\Delta^2 = \frac{\mathbb{E}\left(\rho'(Z)\right)^2}{(L'(0))^2}$.

Asymptotic results

- Question: what happens when $k, N \to \infty$?
- Assume that $\sqrt{k} \cdot g(n) \to 0$ as $N \to \infty$ (if $\mathbb{E}|X-\mu|^{2+\delta} < \infty$, then $k = o\left(N^{\frac{\delta}{1+\delta}}\right)$ suffices).
- $L(z) := \mathbb{E}\rho'\left(z+Z\right)$, where $Z \sim N(0,1)$.
- $\Delta^2 = \frac{\mathbb{E}\left(\rho'(Z)\right)^2}{(L'(0))^2}$.

Theorem (M., 2018)

Under these assumptions,

$$\sqrt{N}\left(\widehat{\mu}^{(k)} - \mu\right) \xrightarrow{d} N\left(0, \Delta^2\sigma^2\right).$$

Asymptotic results

- Question: what happens when $k, N \to \infty$?
- Assume that $\sqrt{k} \cdot g(n) \to 0$ as $N \to \infty$ (if $\mathbb{E}|X - \mu|^{2+\delta} < \infty$, then $k = o\left(N^{\frac{\delta}{1+\delta}}\right)$ suffices).
- $L(z) := \mathbb{E}\rho'(z + Z)$, where $Z \sim N(0,1)$.
- $\Delta^2 = \frac{\mathbb{E}\left(\rho'(Z)\right)^2}{(L'(0))^2}$.

Theorem (M., 2018)

Under these assumptions,

$$\sqrt{N}\left(\widehat{\mu}^{(k)} - \mu\right) \xrightarrow{d} N\left(0, \Delta^2\sigma^2\right).$$

- $\rho(x) = |x| \implies \Delta^2 = \frac{\pi}{2}$.
- $\rho(x) = \begin{cases} z^2/2, & |z| \le M, \\ M|z| - M^2/2, & |z| > M \end{cases} \implies \Delta^2 = \frac{\int_{-M}^{M} x^2 d\Phi(x) + 2M^2(1-\Phi(M))}{(2\Phi(M)-1)^2}$.

  For instance, $\Delta^2 \simeq 1.15$ for $M = 2$ and $\Delta^2 \simeq 1.01$ for $M = 3$.

Uniform deviation bounds

- $X_1, \ldots, X_N$ i.i.d. copies of $X \in S$, $\mathcal{F}$ is a class of functions $f : S \mapsto \mathbb{R}$.

Uniform deviation bounds

- $X_1, \ldots, X_N$ i.i.d. copies of $X \in S$, $\mathcal{F}$ is a class of functions $f : S \mapsto \mathbb{R}$.
- Problem: estimate $\mathbb{E}f(X)$ for all $f \in \mathcal{F}$ (motivated by empirical risk minimization).

Uniform deviation bounds

- $X_1, \ldots, X_N$ i.i.d. copies of $X \in S$, $\mathcal{F}$ is a class of functions $f : S \mapsto \mathbb{R}$.
- Problem: estimate $\mathbb{E}f(X)$ for all $f \in \mathcal{F}$ (motivated by empirical risk minimization).
- $\bar{\mu}_j(f) = \frac{1}{n}\sum_{i \in G_j} f(X_i)$, $\sigma^2(\mathcal{F}) = \sup_{f \in \mathcal{F}} \mathrm{Var}(f(X))$, and

$$\widehat{\mu}^{(k)}(f) := \operatorname{argmin}_{z \in \mathbb{R}} \frac{1}{\sqrt{N}} \sum_{j=1}^{k} \rho\left(\sqrt{n}\,\frac{\bar{\mu}_j(f) - z}{\Delta}\right), \text{ where } \Delta > 0.$$

Uniform deviation bounds

- $X_1, \ldots, X_N$ i.i.d. copies of $X \in S$, $\mathcal{F}$ is a class of functions $f : S \mapsto \mathbb{R}$.
- Problem: estimate $\mathbb{E}f(X)$ for all $f \in \mathcal{F}$ (motivated by empirical risk minimization).
- $\bar{\mu}_j(f) = \frac{1}{n}\sum_{i \in G_j} f(X_i)$, $\sigma^2(\mathcal{F}) = \sup_{f \in \mathcal{F}} \mathrm{Var}(f(X))$, and

$$\widehat{\mu}^{(k)}(f) := \operatorname{argmin}_{z \in \mathbb{R}} \frac{1}{\sqrt{N}} \sum_{j=1}^{k} \rho\left(\sqrt{n}\,\frac{\bar{\mu}_j(f) - z}{\Delta}\right), \text{ where } \Delta > 0.$$

- $g(f; n) := \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\sqrt{n}\frac{\bar{\mu}_j(f) - \mathbb{E}f(X)}{\sqrt{\mathrm{Var}(f(X))}} \le t\right) - \Phi(t) \right|.$

Uniform deviation bounds

$$\sigma^2(\mathcal{F}) = \sup_{f\in\mathcal{F}} \mathrm{Var}(f(X)),\ \widetilde{\Delta} = \max(\Delta, \sigma(\mathcal{F}))$$

Theorem (M., 2018/19)

Assume that ρ' is Lipschitz continuous. Then for all $s > 0$ such that

$$\max\left(\frac{1}{\sqrt{k}\,\Delta}\,\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left(f(X_j)-\mathbb{E}f(X)\right)\right|, \sqrt{\frac{s}{k}} + \sup_{f\in\mathcal{F}} g(f;n)\right) \le \underset{\sim}{c}(\rho),$$

the inequality

$$\sup_{f\in\mathcal{F}}\left|\widehat{\mu}^{(k)}(f) - \mathbb{E}f(X)\right| \le \bar{C}(\rho)\left(\frac{1}{\sqrt{N}}\frac{\widetilde{\Delta}}{\Delta}\,\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left(f(X_j)-\mathbb{E}f(X)\right)\right| + \widetilde{\Delta}\left(\sqrt{\frac{s}{N}} + \sup_{f\in\mathcal{F}} g(f;n)\sqrt{\frac{k}{N}}\right)\right)$$

holds with probability $\ge 1 - 2e^{-s}$.

Estimators of the mean of a random vector

- $X_1, \ldots, X_N$ – i.i.d. copies of a random vector $X \in \mathbb{R}^d$ with mean $\mathbb{E}X = \mu$ and covariance matrix $\mathbb{E}(X - \mu)(X - \mu)^T = \Sigma$.

Estimators of the mean of a random vector

- $X_1, \dots, X_N$ – i.i.d. copies of a random vector $X \in \mathbb{R}^d$ with mean $\mathbb{E}X = \mu$ and covariance matrix $\mathbb{E}(X-\mu)(X-\mu)^T = \Sigma$.
- $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^d$.

Estimators of the mean of a random vector

- $X_1, \ldots, X_N$ – i.i.d. copies of a random vector $X \in \mathbb{R}^d$ with mean $\mathbb{E}X = \mu$ and covariance matrix $\mathbb{E}(X-\mu)(X-\mu)^T = \Sigma$.
- $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^d$.
- Take $\mathcal{F}$ to be the class of linear functionals $\mathcal{F} = \{f_v(x) = \langle v, x\rangle, \ \|v\|_2 = 1\}$.

Estimators of the mean of a random vector

- $X_1, \ldots, X_N$ – i.i.d. copies of a random vector $X \in \mathbb{R}^d$ with mean $\mathbb{E}X = \mu$ and covariance matrix $\mathbb{E}(X-\mu)(X-\mu)^T = \Sigma$.
- $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^d$.
- Take $\mathcal{F}$ to be the class of linear functionals $\mathcal{F} = \{f_v(x) = \langle v, x\rangle, \ \|v\|_2 = 1\}$.
- Then $\sup_{f\in\mathcal{F}} \left|\widehat{\mu}^{(k)}(f) - \mathbb{E}f(X)\right| = \left\|\widehat{\mu}^{(k)} - \mu\right\|_2$, $\sigma(\mathcal{F}) = \sqrt{\lambda_{\max}(\Sigma)}$ and

$$\frac{1}{N}\,\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^{N}\left(f(X_j) - \mathbb{E}f(X)\right)\right| \le \sqrt{\frac{\operatorname{tr}\Sigma}{N}}.$$

Estimators of the mean of a random vector

- $X_1, \ldots, X_N$ – i.i.d. copies of a random vector $X \in \mathbb{R}^d$ with mean $\mathbb{E}X = \mu$ and covariance matrix $\mathbb{E}(X-\mu)(X-\mu)^T = \Sigma$.
- $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^d$.
- Take $\mathcal{F}$ to be the class of linear functionals $\mathcal{F} = \{f_v(x) = \langle v, x\rangle, \ \|v\|_2 = 1\}$.
- Then $\sup_{f\in\mathcal{F}} \left|\widehat{\mu}^{(k)}(f) - \mathbb{E}f(X)\right| = \left\|\widehat{\mu}^{(k)} - \mu\right\|_2$, $\sigma(\mathcal{F}) = \sqrt{\lambda_{\max}(\Sigma)}$ and

$$\frac{1}{N}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^N \left(f(X_j) - \mathbb{E}f(X)\right)\right| \le \sqrt{\frac{\operatorname{tr}\Sigma}{N}}.$$

- Can be used to construct the estimator $\widehat{\mu}^{(k)}$ that satisfies "sub-Gaussian" bound

$$\left\|\widehat{\mu}^{(k)} - \mu\right\|_2 \le \bar{C}(\rho)\left(\sqrt{\frac{\operatorname{tr}\Sigma}{N}} + \sqrt{\lambda_{\max}(\Sigma)}\left(\sqrt{\frac{s}{N}} + \underbrace{\sup_{v:\|v\|_2=1} g(f_v; n)\sqrt{\frac{k}{N}}}_{\text{"bias" of smaller order}}\right)\right)$$

with probability $\ge 1 - 2e^{-s}$, as long as $k \gtrsim \frac{\operatorname{tr}\Sigma}{\lambda_{\max}(\Sigma)}$ and $s \lesssim k$.

- Construction of $\widehat{\mu}^{(k)}$ (previously been used in the papers by O. Catoni, I. Giulini, G. Lugosi, S. Mendelson, E. Joly and R. Oliveira):

- Construction of $\widehat{\mu}^{(k)}$ (previously been used in the papers by O. Catoni, I. Giulini, G. Lugosi, S. Mendelson, E. Joly and R. Oliveira):
- Let v be the unit vector, and define $X_j(v) := \langle v, X_j \rangle$, and $\bar{\mu}_1(v), \ldots, \bar{\mu}_k(v)$ accordingly.

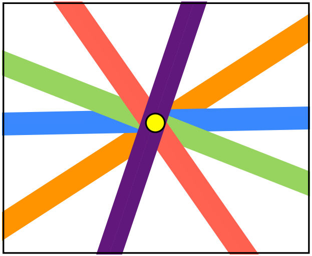
- Construction of $\widehat{\mu}^{(k)}$ (previously been used in the papers by O. Catoni, I. Giulini, G. Lugosi, S. Mendelson, E. Joly and R. Oliveira):
- Let v be the unit vector, and define $X_j(v) := \langle v, X_j \rangle$, and $\bar{\mu}_1(v), \ldots, \bar{\mu}_k(v)$ accordingly.
- "MOM in direction v":

$$\widehat{\mu}^{(k)}(v) := \text{argmin}_{z\in\mathbb{R}} \frac{1}{\sqrt{N}} \sum_{j=1}^{k} \rho\left(\sqrt{n}\, \frac{\bar{\mu}_j(v) - z}{\Delta}\right).$$

- Construction of $\widehat{\mu}^{(k)}$ (previously been used in the papers by O. Catoni, I. Giulini, G. Lugosi, S. Mendelson, E. Joly and R. Oliveira):
- Let v be the unit vector, and define $X_j(v) := \langle v, X_j \rangle$, and $\bar{\mu}_1(v), \ldots, \bar{\mu}_k(v)$ accordingly.
- "MOM in direction v":

$$\widehat{\mu}^{(k)}(v) := \text{argmin}_{z\in\mathbb{R}} \frac{1}{\sqrt{N}} \sum_{j=1}^{k} \rho\left(\sqrt{n}\,\frac{\bar{\mu}_j(v)-z}{\Delta}\right).$$

- $S_v(\varepsilon) := \left\{y \in \mathbb{R}^d : \left|\langle y, v\rangle - \widehat{\mu}^{(k)}(v)\right| \le \varepsilon\right\}, M(\varepsilon) := \bigcap_{v:\|v\|_2=1} S_v(\varepsilon).$



- Construction of $\widehat{\mu}^{(k)}$ (previously been used in the papers by O. Catoni, I. Giulini, G. Lugosi, S. Mendelson, E. Joly and R. Oliveira):
- Let v be the unit vector, and define $X_j(v) := \langle v, X_j \rangle$, and $\bar{\mu}_1(v), \ldots, \bar{\mu}_k(v)$ accordingly.
- "MOM in direction v":
  $$\widehat{\mu}^{(k)}(v) := \text{argmin}_{z \in \mathbb{R}} \frac{1}{\sqrt{N}} \sum_{j=1}^{k} \rho\left(\sqrt{n}\, \frac{\bar{\mu}_j(v) - z}{\Delta}\right).$$
- $S_v(\varepsilon) := \left\{ y \in \mathbb{R}^d : \left| \langle y, v \rangle - \widehat{\mu}^{(k)}(v) \right| \leq \varepsilon \right\}, M(\varepsilon) := \bigcap_{v : \|v\|_2 = 1} S_v(\varepsilon).$
- Finally, let $\varepsilon_* := \inf \{\varepsilon > 0 : M(\varepsilon) \neq \emptyset\}$, and take $\widehat{\mu}^{(k)}$ to be any element in $M(\varepsilon_*)$.

Remark: eliminating dependence on the partition

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

Remark: eliminating dependence on the partition

- $n = \lfloor N/k \rfloor$
- $\mathcal{A}_N^{(n)} = \{J \subset \{1, \ldots, N\} : \ |J| = n\}$ – all subsets of size n; in particular, card$(\mathcal{A}_N^{(n)}) = \binom{N}{n}$

Remark: eliminating dependence on the partition

- $n = \lfloor N/k \rfloor$
- $\mathcal{A}_N^{(n)} = \{J \subset \{1, \dots, N\} : \ |J| = n\}$ – all subsets of size n; in particular, $\mathrm{card}(\mathcal{A}_N^{(n)}) = \binom{N}{n}$
- $\bar{\mu}_J = \text{average}(X_j, \ j \in J)$

Remark: eliminating dependence on the partition

- $n = \lfloor N/k \rfloor$
- $\mathcal{A}_N^{(n)} = \{J \subset \{1, \dots, N\} : \ |J| = n\}$ – all subsets of size n; in particular, $\text{card}(\mathcal{A}_N^{(n)}) = \binom{N}{n}$
- $\bar{\mu}_J = \text{average}(X_j, \ j \in J)$
- $$\widetilde{\mu}_\rho^{(k)} := \operatorname*{argmin}_{z \in \mathbb{R}} \sum_{J \in \mathcal{A}_N^{(n)}} \rho\left(\sqrt{n}\frac{\bar{\mu}_J - z}{\Delta}\right)$$

Remark: eliminating dependence on the partition

- $n = \lfloor N/k \rfloor$
- $\mathcal{A}_N^{(n)} = \{J \subset \{1, \ldots, N\} : \ |J| = n\}$ – all subsets of size n; in particular, $\text{card}(\mathcal{A}_N^{(n)}) = \binom{N}{n}$
- $\bar{\mu}_J = \text{average}(X_j, \ j \in J)$
- $$\widetilde{\mu}_\rho^{(k)} := \underset{z \in \mathbb{R}}{\text{argmin}} \sum_{J \in \mathcal{A}_N^{(n)}} \rho\left(\sqrt{n}\frac{\bar{\mu}_J - z}{\Delta}\right)$$
- Does not depend on random partition and satisfies the same deviation guarantees as $\widehat{\mu}^{(k)}$.

Thank you for listening!