The Median of Means Estimator: Old and New

Stas Minsker

Department of Mathematics, USC

June 2024

New Frontiers in Robust Statistics

[based in part on a joint work with Nate Strawn]

Concentration of measure

- Concentration of measure phenomenon formalizes the idea that

  nice functions of many independent random variables are “essentially constant”

Concentration of measure

- Concentration of measure phenomenon formalizes the idea that

  nice functions of many independent random variables are “essentially constant”

- This idea can serve as a "bridge" between random and deterministic quantities.

Concentration of measure

- Concentration of measure phenomenon formalizes the idea that

  nice functions of many independent random variables are “essentially constant”

- This idea can serve as a "bridge" between random and deterministic quantities.
- Examples include the Gaussian (Borell-TIS) inequality, bounded difference (McDiarmid’s) inequality, Talagrand’s inequality, matrix Bernstein’s inequality, etc.

Concentration of measure

- Concentration of measure phenomenon formalizes the idea that

  nice functions of many independent random variables are "essentially constant"

- This idea can serve as a "bridge" between random and deterministic quantities.
- Examples include the Gaussian (Borell-TIS) inequality, bounded difference (McDiarmid's) inequality, Talagrand's inequality, matrix Bernstein's inequality, etc.
- For example, if $\mathbf{X} = (X_1, \ldots, X_n) \sim N(0, I_n)$ then $\mathbb{E}\|\mathbf{X}\|_2 \in \left[\frac{n}{\sqrt{n+1}}, \sqrt{n}\right]$ and

$$\left| \|\mathbf{X}\|_2 - \mathbb{E}\|\mathbf{X}\|_2 \right| \leq \sqrt{2t}$$

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

Concentration of measure

- For example, if $\mathbf{X} = (X_1, \ldots, X_n) \sim N(0, I_n)$ then $\mathbb{E}\|\mathbf{X}\|_2 \in \left[\frac{n}{\sqrt{n+1}}, \sqrt{n}\right]$ and

$$\left| \|\mathbf{X}\|_2 - \mathbb{E}\|\mathbf{X}\|_2 \right| \leq \sqrt{2t}$$

  with probability at least $1 - e^{-t}$.
- Typically, a.s. boundedness or exponential integrability assumptions are imposed.

What if the random variables of interest have heavy tails?

Concentration of measure

- For example, if $\mathbf{X} = (X_1, \ldots, X_n) \sim N(0, I_n)$ then $\mathbb{E}\|\mathbf{X}\|_2 \in \left[\frac{n}{\sqrt{n+1}}, \sqrt{n}\right]$ and

$$\Big| \|\mathbf{X}\|_2 - \mathbb{E}\|\mathbf{X}\|_2 \Big| \le \sqrt{2t}$$

  with probability at least $1 - e^{-t}$.
- Typically, a.s. boundedness or exponential integrability assumptions are imposed.

What if the random variables of interest have heavy tails?

- For the purpose of this talk, a random variable Z has heavy-tailed distribution if

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

  for some $k > 2$.

Sub-Gaussian mean estimation in $\mathbb{R}$

- $X_1, \ldots, X_N$ – i.i.d. copies of $X \in \mathbb{R}$ such that

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

Sub-Gaussian mean estimation in $\mathbb{R}$

- $X_1, \ldots, X_N$ – i.i.d. copies of $X \in \mathbb{R}$ such that

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

- Goal: construct an estimator $\widehat{\mu}_N$ satisfying

$$\mathbb{P}\left(|\widehat{\mu}_N - \mu| \geq C\sigma\sqrt{\frac{t}{N}} \right) \leq 2e^{-t}$$

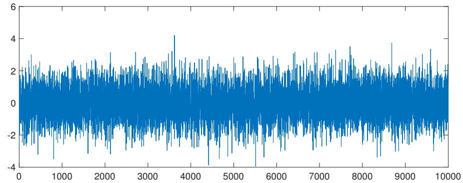
where C is an absolute constant.

Sub-Gaussian mean estimation in $\mathbb{R}$

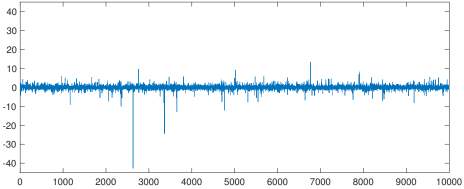
- Goal: construct an estimator $\widehat{\mu}_N$ satisfying

$$\mathbb{P}\left(|\widehat{\mu}_N - \mu| \geq C\sigma\sqrt{\frac{t}{N}}\right) \leq 2e^{-t}$$

where C is an absolute constant.



Standard normal distribution



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

Sub-Gaussian mean estimation in $\mathbb{R}^d$

- $X_1, \ldots, X_N$ – i.i.d. copies of $X \in \mathbb{R}^d$ such that

$$\mathbb{E}X = \mu, \ \mathbb{E}(X-\mu)(X-\mu)^T = \Sigma$$

Sub-Gaussian mean estimation in $\mathbb{R}^d$

- $X_1, \ldots, X_N$ – i.i.d. copies of $X \in \mathbb{R}^d$ such that

$$\mathbb{E}X = \mu, \ \mathbb{E}(X-\mu)(X-\mu)^T = \Sigma$$

- Goal: construct an estimator $\widehat{\mu}_N$ satisfying

$$\mathbb{P}\left(\|\widehat{\mu}_N - \mu\| \geq C_1 \sqrt{\frac{\mathrm{tr}(\Sigma)}{N}} + C_2 \sqrt{\lambda_{\max}(\Sigma)} \sqrt{\frac{t}{N}} \right) \leq e^{-t},$$

where C_1, C_2 are absolute constants, $\|\cdot\|$ - Euclidean norm.

2011 - onwards: large literature on Robustness, both in the Mathematical Statistics and the TCS communities:

- J.-Y. Audibert, A. Minasyan, S. Bahmani, P. Bartlett, V. Brunel, O. Catoni, A. Dalalyan, L. Devroye, G. Depersin, J. Fan, C. Gao, A. Iouditski, Y. Klochkov, J. Kwon, G. Lecué, M. Lerasle, G. Lugosi, S. Mendelson, A. Minasyan, T. Mathieu, M. Ndaoud, R. Oliveira, Z. Rico, A. Tsybakov, I. Giulini, N. Zhivotovskiy.
- Everyone in this audience and beyond..

Heavy tails vs Adversarial Contamination

- Assume that instead of $X_1, \ldots, X_N$, we observe $Y_1, \ldots, Y_N$ where

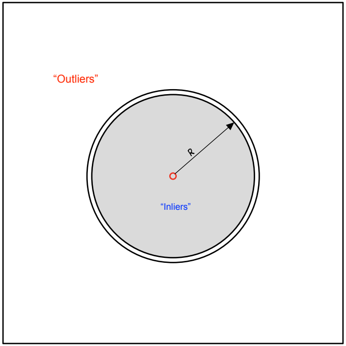
$$Y_j \neq X_j, \; j \in J \text{ for } |J| \leq \varepsilon N$$

Heavy tails vs Adversarial Contamination

- Assume that instead of $X_1, \ldots, X_N$, we observe $Y_1, \ldots, Y_N$ where

$$Y_j \neq X_j, \; j \in J \text{ for } |J| \leq \varepsilon N$$

- Connection to heavy tails (A. Prasad, S. Balakrishnan, P. Ravikumar '19):

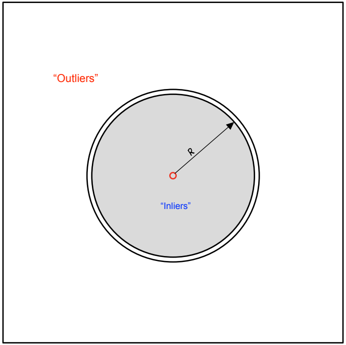


Heavy tails vs Adversarial Contamination

- Assume that instead of $X_1, \ldots, X_N$, we observe $Y_1, \ldots, Y_N$ where

$$Y_j \neq X_j, \ j \in J \text{ for } |J| \leq \varepsilon N$$

- Connection to heavy tails (A. Prasad, S. Balakrishnan, P. Ravikumar '19):



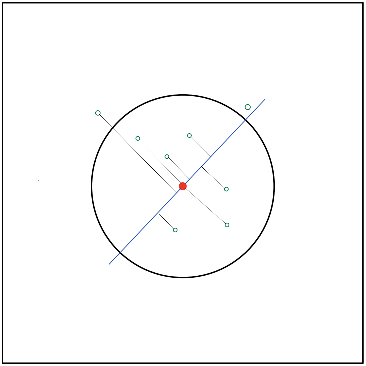
- Works fine in 1d but not in $\mathbb{R}^d$. A better idea: consider each direction separately.

Heavy tails vs Adversarial Contamination

- Assume that instead of $X_1, \ldots, X_N$, we observe $Y_1, \ldots, Y_N$ where

$$Y_j \neq X_j, \ j \in J \text{ for } |J| \leq \varepsilon N$$

- Connection to heavy tails (A. Prasad, S. Balakrishnan, P. Ravikumar '19):
- Works fine in 1d but not in $\mathbb{R}^d$. A better idea: consider each direction separately.

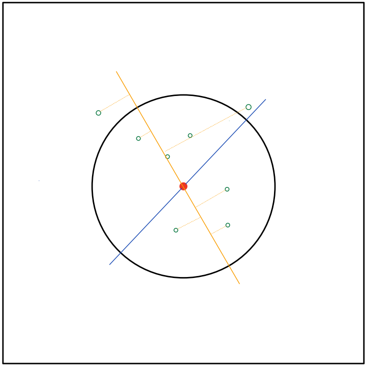


Heavy tails vs Adversarial Contamination

- Assume that instead of $X_1, \ldots, X_N$, we observe $Y_1, \ldots, Y_N$ where

$$Y_j \neq X_j,\ j \in J \text{ for } |J| \leq \varepsilon N$$

- Connection to heavy tails (A. Prasad, S. Balakrishnan, P. Ravikumar '19):
- Works fine in 1d but not in $\mathbb{R}^d$. A better idea: consider each direction separately.



Heavy tails vs Adversarial Contamination

- Assume that instead of $X_1, \ldots, X_N$, we observe $Y_1, \ldots, Y_N$ where

$$Y_j \neq X_j, \; j \in J \text{ for } |J| \leq \varepsilon N$$

- Connection to heavy tails (A. Prasad, S. Balakrishnan, P. Ravikumar '19):
- S. Hopkins, J. Li, F. Zhang '21: both modes of contamination can be solved by "spectral sample reweighing".
- Moreover, the notions of "spectral center" (adversarial) and "combinatorial center" (heavy tails) are equivalent.

Heavy tails vs Adversarial Contamination

- Assume that instead of $X_1, \ldots, X_N$, we observe $Y_1, \ldots, Y_N$ where

$$Y_j \neq X_j, \ j \in J \text{ for } |J| \leq \varepsilon N$$

- Connection to heavy tails (A. Prasad, S. Balakrishnan, P. Ravikumar '19):
- S. Hopkins, J. Li, F. Zhang '21: both modes of contamination can be solved by "spectral sample reweighing".
- Moreover, the notions of "spectral center" (adversarial) and "combinatorial center" (heavy tails) are equivalent.

Sub-Gaussian mean estimation in $\mathbb{R}$

- The Median of Means estimator: early references include *[A. Nemirovski, D. Yudin '83; M. Jerrum, L. Valiant, V. Vazirani '86; N. Alon, Y. Matias, M. Szegedy '96; D. Hsu '10, R. Oliveira, M. Lerasle '11]*
  Split the sample into k "blocks" $G_1, \ldots, G_k$ of size $m \approx N/k$ each

$$\underbrace{\underbrace{\overbrace{X_1, \ldots, X_{|G_1|}}^{G_1}}_{\bar{X}_1 := \frac{1}{|G_1|}\sum_{X_j \in G_1} X_j} \;\cdots\cdots\; \underbrace{\overbrace{X_{N-|G_k|+1}, \ldots, X_N}^{G_k}}_{\bar{X}_k := \frac{1}{|G_k|}\sum_{X_j \in G_k} X_j}}_{\widetilde{\mu}_N := \mathrm{median}(\bar{X}_1, \ldots, \bar{X}_k)}$$

Sub-Gaussian mean estimation in $\mathbb{R}$

- The Median of Means estimator: early references include *[A. Nemirovski, D. Yudin '83; M. Jerrum, L. Valiant, V. Vazirani '86; N. Alon, Y. Matias, M. Szegedy '96; D. Hsu '10, R. Oliveira, M. Lerasle '11]*
  Split the sample into k "blocks" $G_1, \ldots, G_k$ of size $m \approx N/k$ each

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

$$\widetilde{\mu}_N := \text{median}(\bar{X}_1, \ldots, \bar{X}_k)$$

- Then

$$\Pr\left(|\widetilde{\mu}_N - \mu| \geq 7.6 \times \sigma \sqrt{\frac{k}{N}} \right) \leq e^{-k}$$

Sub-Gaussian mean estimation in $\mathbb{R}$

- The Median of Means estimator: early references include *[A. Nemirovski, D. Yudin '83; M. Jerrum, L. Valiant, V. Vazirani '86; N. Alon, Y. Matias, M. Szegedy '96; D. Hsu '10, R. Oliveira, M. Lerasle '11]*
  Split the sample into k "blocks" $G_1, \ldots, G_k$ of size $m \approx N/k$ each
- Then

$$\Pr\left(|\widetilde{\mu}_N - \mu| \geq 7.6 \times \sigma\sqrt{\frac{k}{N}} \right) \leq e^{-k}$$

- Compare to the case of Gaussian distribution:

$$\Pr\left(|\bar{X}_N - \mu| \geq \sqrt{2} \times \sigma\sqrt{\frac{k}{N}} \right) \leq 2e^{-k}$$

Sub-Gaussian mean estimation in $\mathbb{R}$

- The Median of Means estimator: early references include *[A. Nemirovski, D. Yudin '83; M. Jerrum, L. Valiant, V. Vazirani '86; N. Alon, Y. Matias, M. Szegedy '96; D. Hsu '10, R. Oliveira, M. Lerasle '11]*
  Split the sample into k "blocks" $G_1, \ldots, G_k$ of size $m \approx N/k$ each
- Then

$$\Pr\left(|\widetilde{\mu}_N - \mu| \geq 7.6 \times \sigma\sqrt{\frac{k}{N}}\right) \leq e^{-k}$$

- Compare to the case of Gaussian distribution:

$$\Pr\left(|\bar{X}_N - \mu| \geq \sqrt{2} \times \sigma\sqrt{\frac{k}{N}}\right) \leq 2e^{-k}$$

- Is the constant $\sqrt{2} + o(1)$ attainable for heavy-tailed distributions?
- A closely related question of efficiency has been central to mathematical statistics.

Optimal constants

Prior work:

- O. Catoni ‘11; L. Devroye, M. Lerasle, G. Lugosi, R. Oliveira ‘16: $C = \sqrt{2} + o_N(1)$ if an upper bound for the kurtosis is known.

Optimal constants

Prior work:

- O. Catoni '11; L. Devroye, M. Lerasle, G. Lugosi, R. Oliveira '16: $C = \sqrt{2} + o_N(1)$ if an upper bound for the kurtosis is known.
- J. Lee, P. Valiant '22: $C = \sqrt{2} + o_{N,t}(1)$, only finite variance required.

Optimal constants

Prior work:

- O. Catoni '11; L. Devroye, M. Lerasle, G. Lugosi, R. Oliveira '16: $C = \sqrt{2} + o_N(1)$ if an upper bound for the kurtosis is known.
- J. Lee, P. Valiant '22: $C = \sqrt{2} + o_{N,t}(1)$, only finite variance required.
- This talk: $C = \sqrt{2} + o_{P,N}(1)$ for the modified MOM.

Optimal constants

Prior work:

- O. Catoni '11; L. Devroye, M. Lerasle, G. Lugosi, R. Oliveira '16: $C = \sqrt{2} + o_N(1)$ if an upper bound for the kurtosis is known.
- J. Lee, P. Valiant '22: $C = \sqrt{2} + o_{N,t}(1)$, only finite variance required.
- This talk: $C = \sqrt{2} + o_{P,N}(1)$ for the modified MOM.

Optimal constants

Prior work:

- O. Catoni '11; L. Devroye, M. Lerasle, G. Lugosi, R. Oliveira '16: $C = \sqrt{2} + o_N(1)$ if an upper bound for the kurtosis is known.
- J. Lee, P. Valiant '22: $C = \sqrt{2} + o_{N,t}(1)$, only finite variance required.
- This talk: $C = \sqrt{2} + o_{P,N}(1)$ for the modified MOM.

MOM and U-statistics

- Let $\widetilde{\Phi}_m$ be the distribution of $\frac{1}{m}\sum_{j=1}^{m} X_j$.

MOM and U-statistics

- Let $\widetilde{\Phi}_m$ be the distribution of $\frac{1}{m}\sum_{j=1}^{m} X_j$.
- median $\left(\widetilde{\Phi}_m\right)$ minimizes $F(z) = \mathbb{E}\left|\frac{1}{m}\sum_{j=1}^{m} X_j - z\right|$.

MOM and U-statistics

- Let $\widetilde{\Phi}_m$ be the distribution of $\frac{1}{m}\sum_{j=1}^m X_j$.
- median $\left(\widetilde{\Phi}_m\right)$ minimizes $F(z) = \mathbb{E}\left|\frac{1}{m}\sum_{j=1}^m X_j - z\right|$.
- A UMVUE of $F(z)$ is the U-statistic [Halmos, '46, Hoeffding '48, Fraser '54]

$$F_N(z) := \frac{1}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} \left|\bar{X}_J - z\right|$$

where $\mathcal{A}_N^{(m)} = \{J \subset \{1, \ldots, N\} : \ |J| = m\}$ and $\bar{X}_J = \frac{1}{m}\sum_{i \in J} X_i$.

MOM and U-statistics

- Let $\widetilde{\Phi}_m$ be the distribution of $\frac{1}{m}\sum_{j=1}^m X_j$.
- median $\left(\widetilde{\Phi}_m\right)$ minimizes $F(z) = \mathbb{E}\left|\frac{1}{m}\sum_{j=1}^m X_j - z\right|$.
- A UMVUE of $F(z)$ is the U-statistic [Halmos, '46, Hoeffding '48, Fraser '54]

$$F_N(z) := \frac{1}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} \left|\bar{X}_J - z\right|$$

where $\mathcal{A}_N^{(m)} = \{J \subset \{1, \ldots, N\} : \; |J| = m\}$ and $\bar{X}_J = \frac{1}{m}\sum_{i \in J} X_i$.

- Define

$$\boxed{\widehat{\mu}_N := \operatorname*{argmin}_{z \in \mathbb{R}} \frac{1}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} \left|z - \bar{X}_J\right| = \text{median}\left(\bar{X}_J, \; J \in \mathcal{A}_N^{(m)}\right)}$$

Alternatively, $\widehat{\mu}_N$ is the Hodges-Lehmann estimator of order m.

MOM and U-statistics

- Define

$$\boxed{\widehat{\mu}_N := \underset{z\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{\binom{N}{m}} \sum_{J\in\mathcal{A}_N^{(m)}} \left|z - \bar{X}_J\right| = \text{median}\left(\bar{X}_J,\ J \in \mathcal{A}_N^{(m)}\right)}$$

  Alternatively, $\widehat{\mu}_N$ is the Hodges-Lehmann estimator of order m.

- For example, if $N = 4$ and $m = 2$, there will be 6 means:

$$\frac{X_1 + X_2}{2},\ \frac{X_1 + X_3}{2},\ \frac{X_1 + X_4}{2},\ \frac{X_2 + X_3}{2},\ \frac{X_2 + X_4}{2},\ \frac{X_3 + X_4}{2}$$

  versus 2 means for the “standard” MOM: $\frac{X_1+X_2}{2}$, $\frac{X_3+X_4}{2}$.

MOM and U-statistics

- Do we need to include the blocks that are nearly identical?

MOM and U-statistics

- Do we need to include the blocks that are nearly identical?
- Improvement: only leave the blocks of data that are "sufficiently different".

MOM and U-statistics

- Example: sample size $N = 8$, block size $m = 4$, and let

$$Z_1 = \frac{X_1 + X_2}{2},\ Z_2 = \frac{X_3 + X_4}{2},\ Z_3 = \frac{X_5 + X_6}{2},\ Z_4 = \frac{X_7 + X_8}{2}$$

  Now form all averages among the pairs of Z's: we will have 6 means.

MOM and U-statistics

- Example: sample size $N = 8$, block size $m = 4$, and let

$$Z_1 = \frac{X_1 + X_2}{2},\ Z_2 = \frac{X_3 + X_4}{2},\ Z_3 = \frac{X_5 + X_6}{2},\ Z_4 = \frac{X_7 + X_8}{2}$$

  Now form all averages among the pairs of Z's: we will have 6 means.

- Compare to the standard MOM: 2 means, and
  "permutation-invariant" MOM: $\binom{8}{4} = 70$ means.

MOM and U-statistics

- Example: sample size $N = 8$, block size $m = 4$, and let

$$Z_1 = \frac{X_1 + X_2}{2},\ Z_2 = \frac{X_3 + X_4}{2},\ Z_3 = \frac{X_5 + X_6}{2},\ Z_4 = \frac{X_7 + X_8}{2}$$

  Now form all averages among the pairs of Z's: we will have 6 means.

- If m is the size of each "block," it suffices to consider blocks which differ by at least $\frac{m}{\log(m)}$ points.
- Formally, let $n = \frac{N}{m}\lfloor \log(m) \rfloor$, and create a "new sample" $Z_1, \ldots, Z_n$ using mini-batches of size $\ell = m/\lfloor \log(m) \rfloor$:

$$\underbrace{X_1, \ldots, X_{\frac{m}{\lfloor \log(m) \rfloor}}}_{Z_1 := \frac{1}{\ell}\sum_{i=1}^{\ell} X_i} \ldots\ldots \underbrace{X_{N - \frac{m}{\lfloor \log(m) \rfloor} + 1}, \ldots, X_N}_{Z_n := \frac{1}{\ell}\sum_{i=N-\ell+1}^{N} X_i}$$

MOM and U-statistics

- Example: sample size $N = 8$, block size $m = 4$, and let

$$Z_1 = \frac{X_1 + X_2}{2},\ Z_2 = \frac{X_3 + X_4}{2},\ Z_3 = \frac{X_5 + X_6}{2},\ Z_4 = \frac{X_7 + X_8}{2}$$

  Now form all averages among the pairs of Z's: we will have 6 means.

- If m is the size of each "block," it suffices to consider blocks which differ by at least $\frac{m}{\log(m)}$ points.

- Formally, let $n = \frac{N}{m}\lfloor \log(m) \rfloor$, and create a "new sample" $Z_1, \ldots, Z_n$ using mini-batches of size $\ell = m/\lfloor \log(m) \rfloor$:

$$\underbrace{X_1, \ldots, X_{\frac{m}{\lfloor \log(m) \rfloor}}}_{Z_1 := \frac{1}{\ell}\sum_{i=1}^{\ell} X_i} \cdots\cdots \underbrace{X_{N-\frac{m}{\lfloor \log(m) \rfloor}+1}, \ldots, X_N}_{Z_n := \frac{1}{\ell}\sum_{i=N-\ell+1}^{N} X_i}$$

- Define

$$\boxed{\widehat{\mu}'_N := \text{median}\left(\bar{Z}_J,\ J \in \mathcal{A}_n^{(\lfloor \log(m) \rfloor)}\right)}$$

  where $\mathcal{A}_n^{(\ell)} = \{J \subset \{1, \ldots, n\} :\ |J| = \lfloor \log(m) \rfloor\}$ and $\bar{X}_J = \frac{1}{\lfloor \log(m) \rfloor}\sum_{i \in J} Z_i$.

Performance guarantees

Theorem (M. '23)

Assume that $\mathbb{E}\,|(X-\mu)/\sigma|^{2+\varepsilon} < \infty$ *for some* $\varepsilon > 0$. *Then for any* $1 \leq t = o(N/\log^2(N))$ *there exists a version of* $\widehat{\mu}'_N$ *such that*

$$\mathbb{P}\left(|\widehat{\mu}'_N - \mu| \geq (\sqrt{2} + o_{P,N}(1))\sigma\sqrt{\frac{t}{N}}\right) \leq (2 + o_N(1))e^{-t}.$$

Proof ideas

- Problem: understand concentration properties of U-statistics

$$U_{N,m}(h) = \frac{1}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} h\left(X_i,\ i \in J\right)$$

where h is bounded and $m = m(N)$ grows with N.

Variance of U-stiatistics

- Hoeffding's decomposition: $U_{N,m}(h) = \frac{1}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} h(X_i,\ i \in J)$,

$$U_{N,m}(h) - \mathbb{E}U_{N,m}(h) = \underbrace{\frac{m}{N} \sum_{j=1}^{N} \mathbb{E}\Big[h(X_1, \ldots, X_m) \,|\, X_i\Big]}_{\text{Hájek projection}} + \text{Remainder}$$

Variance of U-stiatistics

- Hoeffding's decomposition: $U_{N,m}(h) = \frac{1}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} h\left(X_i,\ i \in J\right)$,

$$U_{N,m}(h) - \mathbb{E}U_{N,m}(h) = \underbrace{\frac{m}{N}\sum_{j=1}^{N} \mathbb{E}\left[h(X_1, \ldots, X_m) \,|\, X_i\right]}_{\text{Hájek projection}} + \text{Remainder}$$

Key challenge: the remainder is a function of random variables with small variance and large sup-norm.

Mean estimation in $\mathbb{R}^d$ (joint with N. Strawn)

- $X_1, \ldots, X_N$ – i.i.d. copies of $X \in \mathbb{R}^d$ such that

$$\mathbb{E}X = \mu,\ \mathbb{E}(X-\mu)(X-\mu)^T = \Sigma$$

Mean estimation in $\mathbb{R}^d$ (joint with N. Strawn)

- $X_1, \ldots, X_N$ – i.i.d. copies of $X \in \mathbb{R}^d$ such that

$$\mathbb{E}X = \mu,\ \mathbb{E}(X-\mu)(X-\mu)^T = \Sigma$$

- Goal: construct an estimator $\widehat{\mu}_N$ satisfying

$$\mathbb{P}\left(\|\widehat{\mu}_N - \mu\| \geq C_1 \sqrt{\frac{\text{tr}(\Sigma)}{N}} + C_2 \sqrt{\lambda_{\max}(\Sigma)} \sqrt{\frac{t}{N}} \right) \leq e^{-t},$$

where C_1, C_2 are absolute constants, $\|\cdot\|$ - Euclidean norm.

Mean estimation in $\mathbb{R}^d$ (joint with N. Strawn)

- $X_1, \ldots, X_N$ – i.i.d. copies of $X \in \mathbb{R}^d$ such that

$$\mathbb{E}X = \mu, \; \mathbb{E}(X-\mu)(X-\mu)^T = \Sigma$$

- Goal: construct an estimator $\widehat{\mu}_N$ satisfying

$$\mathbb{P}\left(\|\widehat{\mu}_N - \mu\| \geq C_1 \sqrt{\frac{\mathrm{tr}(\Sigma)}{N}} + C_2 \sqrt{\lambda_{\max}(\Sigma)} \sqrt{\frac{t}{N}} \right) \leq e^{-t},$$

where C_1, C_2 are absolute constants, $\|\cdot\|$ - Euclidean norm.

- "Geometric" median of means:

$$\widetilde{\mu}_N = \operatorname*{argmin}_{z \in \mathbb{R}^d} \sum_{j=1}^{k} \|z - \bar{X}_j\|$$

Mean estimation in $\mathbb{R}^d$ (joint with N. Strawn)

- Goal: construct an estimator $\widehat{\mu}_N$ satisfying

$$\mathbb{P}\left(\|\widehat{\mu}_N - \mu\| \geq C_1 \sqrt{\frac{\mathrm{tr}(\Sigma)}{N}} + C_2 \sqrt{\lambda_{\max}(\Sigma)} \sqrt{\frac{t}{N}} \right) \leq e^{-t},$$

where C_1, C_2 are absolute constants, $\|\cdot\|$ - Euclidean norm.

- "Geometric" median of means:

$$\widetilde{\mu}_N = \underset{z \in \mathbb{R}^d}{\mathrm{argmin}} \sum_{j=1}^{k} \|z - \bar{X}_j\|$$

- It satisfies, with $k = \lfloor 4t \rfloor + 1$,

$$\mathbb{P}\left(\|\widetilde{\mu}_N - \mu\| \geq 11 \sqrt{\frac{\mathrm{tr}(\Sigma) \cdot t}{N}} \right) \leq 2e^{-t}$$

$\Longrightarrow$ sub-Gaussian deviations when $r(\Sigma) := \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|}$ is small.

Mean estimation in $\mathbb{R}^d$ (joint with N. Strawn)

- "Geometric" median of means:

$$\widetilde{\mu}_N = \operatorname*{argmin}_{z \in \mathbb{R}^d} \sum_{j=1}^{k} \|z - \bar{X}_j\|$$

- It satisfies, with $k = \lfloor 4t \rfloor + 1$,

$$\mathbb{P}\left(\|\widetilde{\mu}_N - \mu\| \geq 11 \sqrt{\frac{\mathrm{tr}(\Sigma) \cdot t}{N}} \right) \leq 2e^{-t}$$

$\Longrightarrow$ sub-Gaussian deviations when $r(\Sigma) := \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|}$ is small.

Mean estimation in $\mathbb{R}^d$ (joint with N. Strawn)

- "Geometric" median of means:

$$\widetilde{\mu}_N = \operatorname*{argmin}_{z\in\mathbb{R}^d} \sum_{j=1}^{k} \|z - \bar{X}_j\|$$

- It satisfies, with $k = \lfloor 4t \rfloor + 1$,

$$\mathbb{P}\left(\|\widetilde{\mu}_N - \mu\| \geq 11\sqrt{\frac{\mathrm{tr}(\Sigma)\cdot t}{N}} \right) \leq 2e^{-t}$$

$\Longrightarrow$ sub-Gaussian deviations when $r(\Sigma) := \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|}$ is small.

- Is it the best possible bound? No: for large classes of distributions P,

$$\mathbb{P}\left(\|\widetilde{\mu}_N - \mu\| \geq C(P)\left(\sqrt{\frac{\mathrm{tr}(\Sigma)}{N}} + \sqrt{\lambda_{\max}(\Sigma)}\sqrt{\frac{t}{N}} \right) \right) \leq e^{-\sqrt{t}}.$$

Improved bounds for the geometric MOM

Let $\widetilde{\Phi}_m$ be the distribution of $\bar{X}_m = \frac{1}{m}\sum_{j=1}^{m} X_j$. Then

$$\widetilde{\mu}_N - \mu = \underbrace{\text{median}\left(\widetilde{\Phi}_m\right) - \mu}_{\text{“bias”}} + \underbrace{\widetilde{\mu}_N - \text{median}\left(\widetilde{\Phi}_m\right)}_{\text{stochastic error}}$$

Improved bounds for the geometric MOM

Let $\widetilde{\Phi}_m$ be the distribution of $\bar{X}_m = \frac{1}{m}\sum_{j=1}^{m} X_j$. Then

$$\widetilde{\mu}_N - \mu = \underbrace{\text{median}\left(\widetilde{\Phi}_m\right) - \mu}_{\text{"bias"}} + \underbrace{\widetilde{\mu}_N - \text{median}\left(\widetilde{\Phi}_m\right)}_{\text{stochastic error}}$$

Theorem (M., N. Strawn)

Assume that Y has absolutely continuous distribution P_Y on a subspace of $\mathbb{R}^d$. Then

$$\|median\,(P_Y) - \mu\| \leq \min\left(\sqrt{tr(\Sigma_Y)}, \sqrt{\|\Sigma_Y\|}\,\frac{\mathbb{E}^{1/2}\,\|Y - median\,(P_Y)\|^{-2}}{\mathbb{E}\,\|Y - median\,(P_Y))\|^{-1}}\right).$$

Improved bounds for the geometric MOM

Let $\widetilde{\Phi}_m$ be the distribution of $\bar{X}_m = \frac{1}{m}\sum_{j=1}^m X_j$. Then

$$\widetilde{\mu}_N - \mu = \underbrace{\text{median}\left(\widetilde{\Phi}_m\right) - \mu}_{\text{“bias”}} + \underbrace{\widetilde{\mu}_N - \text{median}\left(\widetilde{\Phi}_m\right)}_{\text{stochastic error}}$$

Theorem (M., N. Strawn)

Assume that Y has absolutely continuous distribution P_Y on a subspace of $\mathbb{R}^d$. Then

$$\|median\,(P_Y) - \mu\| \leq \min\left(\sqrt{tr(\Sigma_Y)}, \sqrt{\|\Sigma_Y\|}\,\frac{\mathbb{E}^{1/2}\,\|Y - median\,(P_Y)\|^{-2}}{\mathbb{E}\,\|Y - median\,(P_Y))\|^{-1}}\right).$$

Note that

$$\mathbb{E}^{1/2}\,\|Y - \text{median}\,(P_Y)\|^{-2} = \int_0^\infty \underbrace{\mathbb{P}\Big(\|Y - \text{median}\,(P_Y)\,\|^2 \leq t\Big)}_{\text{“small ball” probability}}\frac{dt}{t^2}$$

Equivalence of negative moments of the norm

Lemma (M., N. Strawn)

Assume that Y *has normal distribution* $N(0, \Sigma_Y)$ *such that the effective rank of the covariance matrix* $r(\Sigma_Y) > 10$*. Then*

$$\frac{\mathbb{E}^{1/2} \| Y - \mathit{median}\,(P_Y)\|^{-2}}{\mathbb{E}\, \| Y - \mathit{median}\,(P_Y)\|^{-1}} \leq C$$

for an absolute constant C*.*

Equivalence of negative moments of the norm

Given an absolutely continuous random vector/variable X with density p_X, let

$$M(X) := \|p_X\|_\infty$$

Lemma (S.M., N. Strawn '23)

Assume that $Y \in \mathbb{R}^d$ is given by a linear transformation

$$Y = AZ$$

where $Z = \left(Z^{(1)}, \ldots, Z^{(k)}\right) \in \mathbb{R}^k$ is a random vector with independent coordinates such that $\Sigma_Z = I_k$. Moreover, suppose that $r(\Sigma_Y) \geq 4$. Then

$$\frac{\mathbb{E}^{1/2} \|Y - median\,(P_Y)\|^{-2}}{\mathbb{E}\,\|Y - median\,(P_Y)\|^{-1}} \leq C \max_{j=1,\ldots,k} M(Z^{(j)})$$

for an absolute constant C.

Equivalence of negative moments of the norm

Given an absolutely continuous random vector/variable X with density p_X, let

$$M(X) := \|p_X\|_\infty$$

Lemma (S.M., N. Strawn '23)

Let $Y \in \mathbb{R}^d$, $d \geq 3$ be a random vector with absolutely continuous distribution and covariance matrix Σ_Y. Then

$$\frac{\mathbb{E}^{1/2}\left\|Y - \text{median}\,(P_Y)\right\|^{-2}}{\mathbb{E}\left\|Y - \text{median}\,(P_Y)\right\|^{-1}} \leq C M^{1/d}\left(\Sigma_Y^{-1/2} Y\right)\sqrt{\frac{\sum_{j=1}^{d}\lambda_j}{d\left(\prod_{i=1}^{d}\lambda_i\right)^{1/d}}}$$

for an absolute constant C, where $\lambda_1 \geq \ldots \geq \lambda_d$ are the eigenvalues of Σ_Y.

Equivalence of negative moments of the norm

Given an absolutely continuous random vector/variable X with density p_X, let

$$M(X) := \|p_X\|_\infty$$

Lemma (S.M., N. Strawn '23)

Let $Y \in \mathbb{R}^d$, $d \geq 3$ be a random vector with absolutely continuous distribution and covariance matrix Σ_Y. Then

$$\frac{\mathbb{E}^{1/2}\,\|Y - median\,(P_Y)\|^{-2}}{\mathbb{E}\,\|Y - median\,(P_Y)\|^{-1}} \leq CM^{1/d}\left(\Sigma_Y^{-1/2}Y\right)\sqrt{\frac{\sum_{j=1}^d \lambda_j}{d\left(\prod_{i=1}^d \lambda_i\right)^{1/d}}}$$

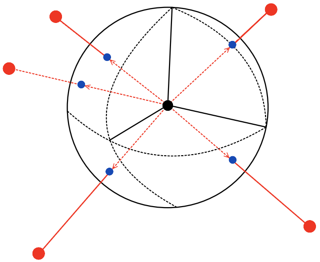
for an absolute constant C, where $\lambda_1 \geq \ldots \geq \lambda_d$ are the eigenvalues of Σ_Y.

- For example, if $\lambda_j = \frac{C}{j^\alpha}$ for $\alpha < 1$, then

$$\frac{\sum_{j=1}^d \lambda_j}{d\left(\prod_{i=1}^d \lambda_i\right)^{1/d}} \leq C(\alpha).$$

- Extensions to “perturbations” of distributions with nice covariance structrures.

Main results



- Stochastic error: key observation is that

$$\left\| \widetilde{\mu}_N - \text{median}\left(\widetilde{\Phi}_m\right) \right\| \lesssim \sqrt{\frac{\text{tr}(\Sigma)k}{N}} \left\| \frac{1}{k} \sum_{j=1}^{k} \frac{\bar{X}_j - m}{\|\bar{X}_j - m\|} \right\|$$

Main results

Theorem (M., N. Strawn)

Assume that $Y \in \mathbb{R}^d$ has "nice" heavy-tailed distribution P. Then

$$\|\widetilde{\mu}_N - \mu\| \leq C_P \left(\sqrt{\frac{tr(\Sigma)}{N}} + \sqrt{\|\Sigma\|}\sqrt{\frac{k}{N}} \right)$$

with probability at least $1 - e^{-\sqrt{k}}$.

Some open questions

- Are there natural classes of heavy-tailed distributions for which the geometric median of means achieves sub-Gaussian performance?
- Can one construct multivariate robust mean estimators with “optimal” constants?