Mean estimation: median-of-means tournaments

Gábor Lugosi

ICREA, Pompeu Fabra University, BGSE

based on joint work with
Luc Devroye (McGill, Montreal)
Matthieu Lerasle (CNRS, Orsay)
Roberto Imbuzeiro Oliveira (IMPA, Rio)
Shahar Mendelson (Technion and ANU)

estimating the mean

Given $X_1, \ldots, X_n$, a real i.i.d. sequence, estimate $\mu = \mathbb{E}X_1$.

estimating the mean

Given $X_1, \ldots, X_n$, a real i.i.d. sequence, estimate $\mu = \mathbb{E}X_1$.

"Obvious" choice: empirical mean

$$\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

estimating the mean

Given $X_1, \ldots, X_n$, a real i.i.d. sequence, estimate $\mu = \mathbb{E}X_1$.

"Obvious" choice: empirical mean

$$\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

By the central limit theorem, if X has a finite variance σ^2,

$$\lim_{n\to\infty} \mathbb{P}\left\{\sqrt{n}\,|\overline{\mu}_n - \mu| > \sigma\sqrt{2\log(2/\delta)}\right\} \le \delta\ .$$

We would like non-asymptotic inequalities of a similar form.

estimating the mean

Given $X_1, \dots, X_n$, a real i.i.d. sequence, estimate $\mu = \mathbb{E}X_1$.

"Obvious" choice: empirical mean

$$\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

By the central limit theorem, if X has a finite variance σ^2,

$$\lim_{n\to\infty} \mathbb{P}\left\{\sqrt{n}\,|\overline{\mu}_n - \mu| > \sigma\sqrt{2\log(2/\delta)}\right\} \le \delta\ .$$

We would like non-asymptotic inequalities of a similar form.

If the distribution is sub-Gaussian,
$\mathbb{E}\exp(\lambda(X-\mu)) \le \exp(\sigma^2\lambda^2/2)$, then with probability at least $1-\delta$,

$$|\overline{\mu}_n - \mu| \le \sigma\sqrt{\frac{2\log(2/\delta)}{n}}.$$

empirical mean–heavy tails

The empirical mean is computationally attractive.

Requires no a priori knowledge and automatically scales with σ.

If the distribution is not sub-Gaussian, we still have Chebyshev's inequality: w.p. $\geq 1 - \delta$,

$$|\overline{\mu}_n - \mu| \leq \sigma \sqrt{\frac{1}{n\delta}} \ .$$

Exponentially weaker bound. Especially hurts when many means are estimated simultaneously.
This is the best one can say. Catoni (2012) shows that for each δ there exists a distribution with variance σ such that

$$\mathbb{P}\left\{ |\overline{\mu}_n - \mu| \geq \sigma \sqrt{\frac{c}{n\delta}} \right\} \geq \delta \ .$$

median of means

A simple estimator is median-of-means. Goes back to Nemirovsky, Yudin (1983), Jerrum, Valiant, and Vazirani (1986), Alon, Matias, and Szegedy (2002).

$$\widehat{\mu}_{MM} \stackrel{\text{def}}{=} \text{median}\left(\frac{1}{m}\sum_{t=1}^{m} X_t, \ldots, \frac{1}{m}\sum_{t=(k-1)m+1}^{km} X_t\right)$$

median of means

A simple estimator is median-of-means. Goes back to Nemirovsky, Yudin (1983), Jerrum, Valiant, and Vazirani (1986), Alon, Matias, and Szegedy (2002).

$$\widehat{\mu}_{MM} \stackrel{\text{def}}{=} \text{median}\left(\frac{1}{m}\sum_{t=1}^{m} X_t, \ldots, \frac{1}{m}\sum_{t=(k-1)m+1}^{km} X_t\right)$$

Lemma

Let $\delta \in (0,1)$, $k = 8\log\delta^{-1}$ *and* $m = \frac{n}{8\log\delta^{-1}}$. *Then with probability at least* $1-\delta$,

$$|\widehat{\mu}_{MM} - \mu| \leq \sigma\sqrt{\frac{32\log(1/\delta)}{n}}$$

proof

By Chebyshev, each mean is within distance $\sigma\sqrt{4/m}$ of μ with probability $3/4$.

The probability that the median is not within distance $\sigma\sqrt{4/m}$ of μ is at most $\mathbb{P}\{\text{Bin}(k, 1/4) > k/2\}$ which is exponentially small in k.

median of means

- Sub-Gaussian deviations.
- Scales automatically with σ.
- Parameters depend on required confidence level δ.
- See Lerasle and Oliveira (2012), Hsu and Sabato (2013), Minsker (2014) for generalizations.
- Also works when the variance is infinite. If $\mathbb{E}\left[|X - \mathbb{E}X|^{1+\alpha}\right] = M$ for some $\alpha \leq 1$, then, with probability at least $1 - \delta$,

$$|\widehat{\mu}_{MM} - \mu| \leq \left(8 \frac{(12M)^{1/\alpha} \ln(1/\delta)}{n}\right)^{\alpha/(1+\alpha)}$$

why sub-Gaussian?

Sub-Gaussian bounds are the best one can hope for when the variance is finite.

In fact, for any $M > 0, \alpha \in (0,1]$, $\delta > 2e^{-n/4}$, and mean estimator $\widehat{\mu}_n$, there exists a distribution $\mathbb{E}\left[|X - \mathbb{E}X|^{1+\alpha}\right] = M$ such that

$$|\widehat{\mu}_n - \mu| \geq \left(\frac{M^{1/\alpha}\ln(1/\delta)}{n}\right)^{\alpha/(1+\alpha)} .$$

Proof: The distributions $P_+(0) = 1 - p, P_+(c) = p$ and $P_-(0) = 1 - p, P_-(-c) = p$ are indistinguishable if all n samples are equal to 0.

why sub-Gaussian?

This shows optimality of the median-of-means estimator for all α.

It also shows that finite variance is necessary even for rate $n^{-1/2}$.

One cannot hope to get anything better than sub-Gaussian tails. Catoni proved that sample mean is optimal for the class of Gaussian distributions.

multiple-δ estimators

Do there exist estimators that are sub-Gaussian simultaneously for all confidence levels?

An estimator is multiple-δ -sub-Gaussian for a class of distributions $\mathcal{P}$ and $\delta_{\min}$ if for all $\delta \in [\delta_{\min}, 1)$, and all distributions in $\mathcal{P}$,

$$|\widehat{\mu}_n - \mu| \leq L\sigma\sqrt{\frac{\log(2/\delta)}{n}}.$$

multiple-δ estimators

Do there exist estimators that are sub-Gaussian simultaneously for all confidence levels?

An estimator is multiple-δ -sub-Gaussian for a class of distributions $\mathcal{P}$ and $\delta_{\min}$ if for all $\delta \in [\delta_{\min}, 1)$, and all distributions in $\mathcal{P}$,

$$|\widehat{\mu}_n - \mu| \le L\sigma\sqrt{\frac{\log(2/\delta)}{n}}.$$

The picture is more complex than before.

known variance

Given $0 < \sigma_1 \leq \sigma_2 < \infty$, define the class

$$\mathcal{P}_2^{[\sigma_1^2, \sigma_2^2]} = \{P : \sigma_1^2 \leq \sigma_P^2 \leq \sigma_2^2.\}$$

Let $R = \sigma_2/\sigma_1$.

known variance

Given $0 < \sigma_1 \leq \sigma_2 < \infty$, define the class

$$\mathcal{P}_2^{[\sigma_1^2, \sigma_2^2]} = \{P : \sigma_1^2 \leq \sigma_P^2 \leq \sigma_2^2.\}$$

Let $R = \sigma_2/\sigma_1$.

- If R is bounded then there exists a multiple-δ -sub-Gaussian estimator with $\delta_{\min} = 4e^{1-n/2}$;
- If R is unbounded then there is no multiple-δ -sub-Gaussian estimate for any L and $\delta_{\min} \to 0$.

A sharp distinction.
The exponentially small value of $\delta_{\min}$ is best possible.

construction of multiple-δ estimator

Reminiscent to Lepski's method of adaptive estimation.

For $k = 1, \ldots, K = \log_2(1/\delta_{min})$, use the median-of-means estimator to construct confidence intervals I_k such that

$$\mathbb{P}\{\mu \notin I_k\} \leq 2^{-k} .$$

(This is where knowledge of σ_2 and boundedness of R is used.)
Define

$$\widehat{k} = \min\left\{ k : \bigcap_{j=k}^{K} I_j \neq \emptyset \right\} .$$

Finally, let

$$\widehat{\mu}_n = \text{mid point of} \bigcap_{j=\widehat{k}}^{K} I_j$$

proof

For any $k = 1, \ldots, K$,

$$\mathbb{P}\{|\widehat{\mu}_n - \mu| > |I_k|\} \leq \mathbb{P}\{\mu \notin \cap_{j=k}^K I_j\}$$

because if $\mu \in \cap_{j=k}^K I_j$, then $\cap_{j=k}^K I_j$ is non-empty and therefore $\widehat{\mu}_n \in \cap_{j=k}^K I_j$.
But

$$\mathbb{P}\{\mu \notin \cap_{j=k}^K I_j\} \leq \sum_{j=k}^K \mathbb{P}\{\mu \notin I_j\} \leq 2^{1-k}$$

higher moments

For $\eta \geq 1$ and $\alpha \in (2, 3]$, define

$$\mathcal{P}_{\alpha,\eta} = \{P \, : \, \mathbb{E}|X - \mu|^{\alpha} \leq (\eta\,\sigma)^{\alpha}\} \, .$$

Then for some $C = C(\alpha, \eta)$ there exists a multiple-δ estimator with a constant L and $\delta_{\min} = e^{-n/C}$ for all sufficiently large n.

k-regular distributions

This follows from a more general result:
Define

$$p_-(j) = \mathbb{P}\left\{\sum_{i=1}^{j} X_i \le j\mu\right\} \quad \text{and} \quad p_+(j) = \mathbb{P}\left\{\sum_{i=1}^{j} X_i \ge j\mu\right\} .$$

A distribution is k-regular if

$$\forall j \ge k,\ \min(p_+(j), p_-(j)) \ge 1/3.$$

For this class there exists a multiple-δ estimator with a constant L and $\delta_{\min} = e^{-n/k}$ for all n.

multivariate distributions

Let $\boldsymbol{X}$ be a random vector taking values in $\mathbb{R}^{\boldsymbol{d}}$ with mean $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{X}$ and covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\boldsymbol{T}}$.

Given an i.i.d. sample $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$, we want to estimate $\boldsymbol{\mu}$ that has sub-Gaussian performance.

multivariate distributions

Let X be a random vector taking values in $\mathbb{R}^d$ with mean $\mu = \mathbb{E}X$ and covariance matrix $\Sigma = \mathbb{E}(X - \mu)(X - \mu)^T$.

Given an i.i.d. sample $X_1, \ldots, X_n$, we want to estimate μ that has sub-Gaussian performance.

What is sub-Gaussian?

If X has a multivariate Gaussian distribution, the sample mean $\overline{\mu}_n = (1/n)\sum_{i=1}^n X_1$ satisfies, with probability at least $1 - \delta$,

$$\|\overline{\mu}_n - \mu\| \leq \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{2\lambda_{\max}\log(1/\delta)}{n}} \ ,$$

Can one construct mean estimators with similar performance for a large class of distributions?

high-dimensional median of means

An obvious idea is to use median of means.

Notions of median:

- Coordinate-wise median.
- Geometric median: $\mathrm{argmin}_{y \in \mathbb{R}^d} \sum_{i=1}^{n} \|y - x_i\|$.
- Center of smallest ball containing at least half of the x_i.
- Tukey median.
- ...a new notion introduced here.

coordinate-wise median of means

Coordinate-wise median of means yields the bound:

$$\|\widehat{\mu}_{MM} - \mu\| \leq K\sqrt{\frac{\mathrm{Tr}(\Sigma)\log(d/\delta)}{n}}\ .$$

We can do better.

smallest-ball median

If $\widehat{\mu}_{MM}$ is the center of the smallest ball containing at least half of the block means $Y_j = \frac{1}{m}\sum_{i \in B_j} X_i$, then with probability at least $1-\delta$,

$$\|\widehat{\mu}_{MM} - \mu\| \leq K\sqrt{\frac{\mathrm{Tr}(\Sigma)\log(1/\delta)}{n}} .$$

No further assumption or knowledge of the distribution is required.

Almost sub-Gaussian but not quite.

Dimension free.

Computationally hard.

multivariate median of means

Hsu and Sabato (2013), Minsker (2015) consider geometric median-of-means:

$$\widehat{\mu}_{MM} = \underset{y \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{k=1}^{k} \|y - Y_j\| .$$

Minsker proves that, with probability at least $1 - \delta$,

$$\|\widehat{\mu}_{MM} - \mu\| \leq K \sqrt{\frac{\mathrm{Tr}(\Sigma) \log(1/\delta)}{n}} .$$

Computationally feasible, dimension free.

Still not sub-Gaussian.

median-of-means tournament

We propose a new estimator with a purely sub-Gaussian performance, without further conditions.

The mean μ is the minimizer of $f(x) = \mathbb{E}\|X - x\|^2$.

For any pair $a, b \in \mathbb{R}^d$, we try to guess whether $f(a) < f(b)$ and set up a "tournament".

Partition the data points into k blocks of size $m = n/k$.

We say that a defeats b if

$$\frac{1}{m}\sum_{i \in B_j} \|X_i - a\|^2 < \frac{1}{m}\sum_{i \in B_j} \|X_i - b\|^2$$

on more than $k/2$ blocks B_j.

median-of-means tournament

Within each block compute

$$Y_j = \frac{1}{m} \sum_{i \in B_j} X_i \ .$$

Then a defeats b if

$$\|Y_j - a\| < \|Y_j - b\|$$

on more than $k/2$ blocks B_j.

Lemma. Let $k = \lceil 200 \log(2/\delta) \rceil$. With probability at least $1 - \delta$, μ defeats all $b \in \mathbb{R}^d$ such that $\|b - \mu\| \geq r$, where

$$r = \max \left(800 \left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}, 240 \sqrt{\frac{\lambda_{\max} \log(2/\delta)}{n}} \right) \right) \ .$$

sub-gaussian estimate

For each $a \in \mathbb{R}^d$, define the set

$$S_a = \left\{ x \in \mathbb{R}^d : \text{such that } x \text{ defeats } a \right\}$$

Now define the mean estimator as

$$\widehat{\mu}_n \in \operatorname*{argmin}_{a \in \mathbb{R}^d} radius(S_a) \ .$$

By the lemma, w.p. $\geq 1 - \delta$,

$$radius(S_{\widehat{\mu}_n}) \leq radius(S_\mu) \leq r$$

and therefore

$$\|\widehat{\mu}_n - \mu\| \leq r \ .$$

sub-gaussian performance

Theorem. Let $k = \lceil 200 \log(2/\delta) \rceil$. Then, with probability at least $1-\delta$,

$$\|\widehat{\mu}_n - \mu\| \le r$$

where

$$r = \max\left(800\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{\max}\log(2/\delta)}{n}}\right)\right) .$$

- No other condition other than existence of Σ.
- "Infinite-dimensional" inequality: the same holds in Hilbert spaces.
- The constants are explicit but sub-optimal.

proof of lemma: sketch

Let $\overline{X} = X - \mu$ and $v = b - \mu$. Then μ defeats b if

$$-\frac{1}{m}\sum_{i \in B_j} \langle \overline{X}_i, v \rangle + \|v\|^2 > 0$$

on the majority of blocks B_j. We need to prove that this holds for all v with $\|v\| = r$.

Step 1: For a fixed v, by Chebyshev, with probability at least $9/10$,

$$\left| \frac{1}{m}\sum_{i \in B_j} \langle \overline{X}_i, v \rangle \right| \leq \sqrt{10}\|v\|\sqrt{\frac{\lambda_{\max}}{m}} \leq r^2/2$$

So by a binomial tail estimate, with probability at least $1 - \exp(-k/50)$, this holds on at least $8/10$ of the blocks B_j.

proof sketch

Step 2: Now we take a minimal ϵ cover the set $r \cdot S^{d-1}$ with respect to the norm $\langle v, \Sigma v \rangle^{1/2}$.

This set has $< e^{k/100}$ points if

$$\epsilon = 5r \left(\frac{1}{k} \mathrm{Tr}(\Sigma) \right)^{1/2} ,$$

so we can use the union bound over this ϵ-net.

Step 3: To extend to all points in $r \cdot S^{d-1}$, we need that, with probability at least $1 - \exp(-k/200)$,

$$\sup_{x \in r \cdot S^{d-1}} \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}_{\{|\frac{1}{m} \sum_{i \in B_j} \langle \overline{X}_i, x - v_x \rangle| \geq r^2/2\}} \leq \frac{1}{10} .$$

This may be proved by standard techniques of empirical processes.

algorithmic challenge

Computing the proposed estimator is nontrivial.

Sam Hopkins (2018) gives a semidefinite relaxation of the estimator that can be computed in polynomial time $\boldsymbol{O(nd + (dk)^8)}$.

Catoni and Giulini (2017) and Lecué and Lerasle (2017) define alternative estimates.

general norms

So far we measured accuracy with respect to the Euclidean norm.

Let $X_1, \ldots, X_n$ be i.i.d. in $\mathbb{R}^d$ with mean μ, covariance matrix Σ, and let $\|\cdot\|$ be an arbitrary norm.

What is the best possible accuracy/confidence tade-off? For guidance, we turn to the empirical mean.

empirical mean

For constant "confidence" δ, the empirical mean has accuracy

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right\| \lesssim \frac{\mathbb{E}\|G\|}{\sqrt{n}} ,$$

where $G \sim \mathcal{N}(0, \Sigma)$. When the distribution is sub-Gaussian, for small δ, the empirical mean has accuracy η such that

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right\| \geq \eta\right) \leq \delta .$$

By standard arguments,

$$\eta \leq \frac{C}{\sqrt{n}}\left(\mathbb{E}\|G\| + \sqrt{\log(1/\delta)} \sup_{x^* \in \mathcal{B}^\circ}\left(\mathbb{E}(x^*(X-\mu))^2\right)^{1/2}\right) ,$$

where $\mathcal{B}^\circ$ is the unit ball of the dual of $\|\cdot\|$.

sub-gaussian performance

Question: Is there a mean estimator $\widehat{\mu}_n$ such that, for all distributions with a second moment, with probability $1-\delta$,

$$\|\widehat{\mu}_n - \mu\| \leq \frac{C}{\sqrt{n}}\left(\mathbb{E}\|G\| + \sqrt{\log\frac{1}{\delta}} \sup_{x^* \in \mathcal{B}^\circ}\left(\mathbb{E}(x^*(X-\mu))^2\right)^{1/2}\right) ?$$

Note: in the Euclidean case this coincides with our "sub-Gaussian" notion.

sub-gaussian performance

Question: Is there a mean estimator $\widehat{\mu}_n$ such that, for all distributions with a second moment, with probability $1-\delta$,

$$\|\widehat{\mu}_n - \mu\| \leq \frac{C}{\sqrt{n}}\left(\mathbb{E}\|G\| + \sqrt{\log\frac{1}{\delta}}\sup_{x^*\in\mathcal{B}^\circ}\left(\mathbb{E}(x^*(X-\mu))^2\right)^{1/2}\right) \ ?$$

Note: in the Euclidean case this coincides with our "sub-Gaussian" notion.

Answer: yes.

estimator

- Set $\epsilon > 0$.
- Let $k = \log(2/\delta)$ and split the sample $(X_i)_{i=1}^n$ to k blocks I_j, each of cardinality $m = n/k$. Set $Y_j = \frac{1}{m}\sum_{i \in I_j} X_i$.
- Let T be the set of extreme points of the dual unit ball $\mathcal{B}^\circ$. For every $x^* \in T$ set

$$S_{x^*} = \left\{ y \in \mathbb{R}^d : |x^*(Y_j) - x^*(y)| \le \epsilon \text{ for more than } \frac{k}{2} \text{ blocks} \right\}.$$

- Set $\mathbb{S}(\epsilon) = \bigcap_{x^* \in T} S_{x^*}$ and select $\widehat{\mu}_N(\epsilon, \delta)$ to be any point in $\mathbb{S}(\epsilon)$.

lower bunds

The term

$$\sqrt{\frac{\log(1/\delta)}{n}} \sup_{x^* \in \mathcal{B}^\circ} \left(\mathbb{E}(x^*(X-\mu))^2\right)^{1/2}$$

is necessary even if X is Gaussian. For any estimator $\widehat{\psi}_N$ and any $x^* \in \mathcal{B}^\circ$,

$$\|\widehat{\psi}_N - \mu\| \geq |x^*(\widehat{\psi}_N) - x^*(\mu)| \; .$$

For any fixed $x^* \in \mathcal{B}^\circ$, $x^*(X)$ is real-valued Gaussian. For any mean estimator, the accuracy is at least $n^{-1/2}\sigma\sqrt{\log(2/\delta)}$, and here $\sigma^2 = \mathbb{E}(x^*(X-\mu))^2$.

lower bunds

The term

$$\frac{\|G\|}{\sqrt{n}}$$

is “essentially” necessary. This term appears by bounding the covering numbers of $\mathcal{B}^\circ$ using Sudakov’s inequality. Whenever this step is sharp, there is no estimator that has a better accuracy than $\|G\|/\sqrt{n}$.

references

G. Lugosi and S. Mendelson.
Sub-Gaussian estimators of the mean of a random vector.
Annals of Statistics, to appear.

G. Lugosi and S. Mendelson.
Risk minimization by median-of-means tournaments.
Journal of the European Mathematical Society, to appear.

G. Lugosi and S. Mendelson.
Near-optimal mean estimators with respect to general norms.
preprint, 2018.

L. Devroye, M. Lerasle, G. Lugosi, and R. Imbuzeiro Oliveira.
Sub-Gaussian mean estimators.
Annals of Statistics, 2016.

references

C. Brownlees, E. Joly, and G. Lugosi.
Empirical risk minimization for heavy-tailed losses.
Annals of Statistics, 43:2507–2536, 2015.

E. Joly, and G. Lugosi.
Robust estimation of U-statistics.
Stochastic Processes and their Applications, to appear, 2015.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi.
Bandits with heavy tail.
IEEE Transactions on Information Theory, 59:7711-7717, 2013.