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, Nice)
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,

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

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

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.

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).

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

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* $1-\delta$,

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

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 $\mathcal{P}$,

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

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 $\mathcal{P}$,

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

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

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

(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 $k = 1, \ldots, K$,

$$\mathbb{P}\{|\widehat{\mu}_n - \mu| > |I_k|\} \le \mathbb{P}\{\exists j \ge k : \mu \notin I_j\}$$

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

$$\mathbb{P}\{\exists j \ge k : \mu \notin I_j\} \le \sum_{j=k}^{K} \mathbb{P}\{\mu \notin I_j\} \le 2^{1-k}$$

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 $\boldsymbol{X_1, \ldots, X_n}$, 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 $1 - \delta$,

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

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

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.

multivariate median of means

Hsu and Sabato (2013), Minsker (2015) extended the median-of-means estimate.

Minsker proposes an analogous estimate that uses the multivariate median

$$\mathrm{Med}(x_1, \ldots, x_N) = \underset{y \in \mathbb{R}^d}{argmin} \sum_{i=1}^{N} \|y - x_i\| .$$

For this estimate, 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.

Computationally feasible.

Almost sub-Gaussian but not quite.

Dimension free.

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 $f(x) = \mathbb{E}\|X - \mu\|^2$.

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

$$\widehat{\mu}_N \in \underset{a \in \mathbb{R}^d}{\operatorname{argmin}} \operatorname{radius}(S_a) \ .$$

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

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

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 $9/10$,

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

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 an interesting open problem.

Coordinate descent does not quite do the job—it only guarantees $\|\widehat{\mu}_n - \mu\|_\infty \leq r$.

regression function estimation

Consider the standard statistical supervised learning problem under the squared loss.

Let (X, Y) take values in $\mathcal{X} \times \mathbb{R}$.

The goal is to predict Y, upon observing X, by $f(X)$ for some $f : \mathcal{X} \to \mathbb{R}$.

We measure the quality of f by the risk

$$\mathbb{E}(f(X) - Y)^2 .$$

We have access to a sample $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$.

We choose $\widehat{f}_n$ from a fixed class of functions $\mathcal{F}$. The best function is

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2 .$$

regression function estimation

We measure performance by either the mean squared error

$$\|\widehat{f}_n - f^*\|_{L_2}^2 = \mathbb{E}\big((\widehat{f}_n(X) - f^*(X))^2 | \mathcal{D}_n\big)$$

or by the excess risk

$$R(\widehat{f}_n) = \mathbb{E}\big((\widehat{f}_n(X) - Y)^2 | \mathcal{D}_n\big) - \mathbb{E}(f^*(X) - Y)^2 \ .$$

regression function estimation

We measure performance by either the mean squared error

$$\|\widehat{f}_n - f^*\|_{L_2}^2 = \mathbb{E}\big((\widehat{f}_n(X) - f^*(X))^2 | \mathcal{D}_n\big)$$

or by the excess risk

$$R(\widehat{f}_n) = \mathbb{E}\big((\widehat{f}_n(X) - Y)^2 | \mathcal{D}_n\big) - \mathbb{E}(f^*(X) - Y)^2 \ .$$

A procedure achieves accuracy r with confidence $1 - \delta$ if

$$\mathbb{P}\left(\|\widehat{f}_n - f^*\|_{L_2} \leq r\right) \geq 1 - \delta \ .$$

High accuracy and high confidence are conflicting requirements.

The accuracy edge is the smallest achievable accuracy with confidence $1 - \delta = 3/4$.

A quest with a long history has been to understand the tradeoff.

empirical risk minimization

The standard learning procedure is empirical risk minimization (ERM):

$$\widehat{f}_n = \underset{f \in \mathcal{F}}{\mathrm{argmin}} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 \ .$$

ERM achieves near-optimal accuracy/confidence tradeoff for well-behaved distributions.

The performance of ERM is now well understood.

It works well if both Y and $f(X)$ have sub-Gaussian tails (for all $f \in \mathcal{F}$).

four complexity parameters

The performance of ERM depends on the intricate interplay between the geometry of $\mathcal{F}$ and the distribution of (X, Y).
We assume that $\mathcal{F}$ is convex.

Let $\mathcal{F}_{h,r} = \{f - h : f \in \mathcal{F},\ \|f - h\|_{L_2} \leq r\}$ and let $\mathcal{M}(\mathcal{F}_{h,r}, \epsilon)$ be the ϵ-packing numbers.
For $\kappa, \eta > 0$, set

$$\lambda_{\mathbb{Q}}(\kappa, \eta) = \sup_{h \in \mathcal{F}} \inf\{r : \log \mathcal{M}(\mathcal{F}_{h,r}, \eta r) \leq \kappa^2 n\} \ .$$

Similarly, let

$$\lambda_{\mathbb{M}}(\kappa, \eta) = \sup_{h \in \mathcal{F}} \inf\{r : \log \mathcal{M}(\mathcal{F}_{h,r}, \eta r) \leq \kappa^2 n r^2\}$$

four complexity parameters

$$r_E(\kappa) = \sup_{h \in \mathcal{F}} \inf \left\{ r : \mathbb{E} \sup_{u \in \mathcal{F}_{h,r}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i u(X_i) \right| \le \kappa \sqrt{n} r \right\},$$

Finally, let

$$\begin{aligned} &\bar{r}_{\mathbb{M}}(\kappa, h) \\ &\quad = \inf \left\{ r : \mathbb{E} \sup_{u \in \mathcal{F}_{h,r}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i u(X_i) \cdot (h(X_i) - Y_i) \right| \le \kappa \sqrt{n} r^2 \right\}. \end{aligned}$$

and

$$\widetilde{r}_{\mathbb{M}}(\kappa, \sigma) = \sup_{h \in \mathcal{F}_Y^{(\sigma)}} \bar{r}_{\mathbb{M}}(\kappa, h)$$

where $\mathcal{F}_Y^{(\sigma)} = \{ f \in \mathcal{F} : \|f(X) - Y\|_{L_2} \le \sigma \}$.

accuracy edge

Suppose $\|Y - f^*(X)\|_{L_2} \leq \sigma$ for a known constant $\sigma > 0$.
Introduce the "complexity"

$$r^* = \max\{\lambda_{\mathbb{Q}}(c_1, c_2), \lambda_{\mathbb{M}}(c_1/\sigma, c_2), r_E(c_1), \widetilde{r}_{\mathbb{M}}(c_1, \sigma)\} \ .$$

Mendelson (2016) proved that r^* is an upper bound for the accuracy edge (under a "small-ball" assumption).

linear regression–an example

Let $\mathcal{F} = \{\langle t, \cdot \rangle : t \in \mathbb{R}^d\}$ be the class of linear functionals.

Let X be an isotropic random vector in $\mathbb{R}^d$ such that $\| \langle X, t \rangle \|_{L_4} \leq L \| \langle X, t \rangle \|_{L_2}$.

Suppose $Y = \langle t_0, X \rangle + W$ for some $t_0 \in \mathbb{R}^d$ and symmetric independent noise W with variance σ^2.

linear regression

Given n independent samples (X_i, Y_i), least-squares regression (ERM) finds $\widehat{t}_n$ such that

$$\|\widehat{t}_n - t\| \leq c\frac{\sigma}{\delta}\sqrt{\frac{d}{n}}$$

with probability $1 - \delta - e^{-cd}$.

Note the weak accuracy/confidence tradeoff.

Lecué and Mendelson (2016) show that this is essentially optimal.

However, if everything is sub-Gaussian, one has

$$\|\widehat{t}_n - t\| \leq c\sigma\sqrt{\frac{d}{n}}$$

with probability $1 - e^{-cd}$.

We introduce a procedure that achieves the same performance as sub-Gaussian ERM but under the general fourth-moment condition.

median-of-means tournament

A natural idea is to replace ERM by minimization of the median-of-means estimate of the risk $\mathbb{E}(f(X) - Y)^2$.

Difficult to analyze—may be suboptimal.

median-of-means tournament

A natural idea is to replace ERM by minimization of the median-of-means estimate of the risk $\mathbb{E}(f(X) - Y)^2$.

Difficult to analyze—may be suboptimal.

Instead, we run a median-of-means tournament.

The idea is that, based on a median-of-means estimate of the difference

$$\mathbb{E}(f(X) - Y)^2 - \mathbb{E}(h(X) - Y)^2 ,$$

we can have a good guess if f or h has a smaller risk.

median-of-means tournament

To make the idea work, we design a (two- or) three-step procedure.

Each step uses an independent sample so before starting we split the data into (two or) three equal parts.

The procedure has a parameter $r > 0$, the desired accuracy level.

The main steps of the procedure are:

- Distance referee
- Elimination phase
- Champions league

step 1: the distance referee

For each pair $f, h \in \mathcal{F}$, one may use define a median-of-means estimate $\Phi_n(f, h)$ using $(|f(X_i) - h(X_i)|)_{i=1}^n$ such that, with "high probability", for all $\Phi_n(f, h)$,

$$\text{if} \quad \Phi_n(f, h) \geq \beta r \quad \text{then} \quad \|f - h\|_{L_2} \geq r$$

and

$$\text{if} \quad \Phi_n(f, h) < \beta r \quad \text{then} \quad \|f - h\|_{L_2} < \alpha r$$

for some constants α, β.

Matches are only allowed between $f, h \in \mathcal{F}$ if $\Phi_n(f, h) \geq \beta r$.

step 2: elimination phase

For any pair $f, h \in \mathcal{F}$, if the distance referee allows a match, calculate the median-of-means estimate based on the samples

$$(f(X_i) - Y_i)^2 - (h(X_i) - Y_i)^2 \ .$$

if the estimate is negative, f wins the match otherwise h wins.

$f \in \mathcal{F}$ is a champion if it wins all its matches. Let $\mathcal{H}$ be the set of all champions.

If one only cares about the mean squared error $\|\widehat{f}_n - f^*\|_{L_2}$, then one may select any champion $\widehat{f}_n \in \mathcal{H}$.

One may show that, with "high probability", $\mathcal{H}$ contains f^* and possibly other functions within distance $O(r)$ of f^*.

If the excess risk also matters, all champions in $\mathcal{H}$ advance to the Champions League for the playoffs.

step 3: Champions League

To select a champion with a small excess risk, we use the simple fact that, for any $f \in \mathcal{F}$,

$$\mathbb{E}(f(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2 \\ \leq -2\mathbb{E}(f^*(X) - f(X))(f(X) - Y) .$$

The Champions League winner is selected based on median-of-means estimates of $\mathbb{E}(h(X) - f(X))(f(X) - Y)$ for all pairs $f, h \in \mathcal{F}$.

result

Suppose that $\mathcal{F}$ is a convex class of functions and

- for every $f, h \in \mathcal{F}$, $\|f - h\|_{L_4} \leq L\|f - h\|_{L_2}$;
- for every $f \in \mathcal{F}$, $\|f - Y\|_{L_4} \leq L\|f - Y\|_{L_2}$;

Then the median-of-means tournament achieves an essentially optimal accuracy/confidence tradeoff.

For any $r > r^*$, with probability at least

$$1 - \exp\left(-c_0 n \min\{1, \sigma^{-2} r^2\}\right) ,$$

$$\|\widehat{f} - f^*\|_{L_2} \leq cr$$

and

$$\mathbb{E}\left((\widehat{f}(X) - Y)^2 | \mathcal{D}_n\right) \leq \mathbb{E}(f^*(X) - Y)^2 + (cr)^2 .$$

linear regression

Recall the example $\mathcal{F} = \{\langle t, \cdot \rangle : t \in \mathbb{R}^d\}$ with X isotropic such that $\| \langle X, t \rangle \|_{L_4} \le L \| \langle X, t \rangle \|_{L_2}$ and $Y = \langle t_0, X \rangle + W$.
We obtain

$$\|\widehat{t}_n - t\| \le c\sigma \sqrt{\frac{d}{n}}$$

with probability $1 - e^{-cd}$ and also

$$\mathbb{E}\big((\widehat{f}(X) - Y)^2 | \mathcal{D}_n\big) - \mathbb{E}(f^*(X) - Y)^2 \le c\sigma^2 \frac{d}{n} \ .$$

algorithmic challenge

Find an algorithmically efficient version of the median-of-means tournament.

references

G. Lugosi and S. Mendelson.
Sub-Gaussian estimators of the mean of a random vector.
submitted, 2017.

G. Lugosi and S. Mendelson.
Risk minimization by median-of-means tournaments.
submitted, 2016.

E. Joly, and G. Lugosi, and R. Imbuzeiro Oliveira.
On the estimation of the mean of a random vector.
Electronic Journal of Statistics, 2017.

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.