Course: High Dimensional Probability: applications to statistical learning and data science.

Text(s): "High-Dimensional Probability" Roman Vershynin
"High-Dimensional Statistics" Martin J. Wainwright

Geography: Let n be sample size
d is the dimension of data

- Classical Asymptotics
  $n \nearrow +\infty$, d fixed
  Ex: Laws of large numbers
  Central limit theorem
- High dimensional Asymptotics
  $(n,d) \nearrow +\infty$ while $\psi(n,d) \to \alpha < \infty$,
  where ψ is a "scaling function"
  E.g.: $\psi(n,d) = \frac{d}{n}$
- Non-asymptotic Bounds ⇐ This class
  (n,d) is fixed. Probability of successful estimation/inference depends on (n,d)

See (W:1) for examples of the three viewpoints.

Lecture 1: Review (V:1)

Let X be a random variable on Probability space

Expectation and Variance

$$\mathbb{E}X, \quad \mathrm{Var}(X) = \mathbb{E}(X-\mathbb{E}X)^2 = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

Moment Generating Function:

$$M_X(t) = \mathbb{E}e^{tX}, \quad t \in \mathbb{R}$$

L^p-norm $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$, $p \in (0,\infty)$

Banach Space:

$$L^p = \{X : \|X\|_p < \infty\}$$

Remark: L^2 is a Hilbert space

$$\langle X, Y\rangle_2 = \mathbb{E}[XY], \quad \|X\|_2 = \sqrt{\langle X, X\rangle} = \sqrt{\mathbb{E}[X^2]}$$

Then

$$\|X - \mathbb{E}X\|_2 = \sqrt{\mathrm{Var}(X)} \quad \text{and}$$

$$\mathrm{cov}(X,Y) := \mathbb{E}(X-\mathbb{E}X)(Y-\mathbb{E}Y)$$
$$= \langle X-\mathbb{E}X, Y-\mathbb{E}Y\rangle_2$$

Limit Theorems

Thm (Strong Law of Large Numbers)
Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}X_i = \mu$. Then $S_n = \sum_{i=1}^{n} X_i$ satisfies

$$\frac{1}{n} \cdot S_n \to \mu \quad \text{almost surely}$$

$$\left[\lim_{n\to\infty} \frac{S_n}{n} = \mu \text{ w.p. } 1\right]$$

Intuition: $\operatorname{Var}(S_n) = \dfrac{\operatorname{Var}(X_i)}{n}$

Thm (Central Limit Theorem)
Let $X_1, X_2, \ldots$ be i.i.d. with

$$\mathbb{E}X_i = \mu, \quad \operatorname{Var} X_i = \sigma^2$$

Define $$Z_n = \frac{S_n - \mathbb{E}S_n}{\sqrt{\operatorname{Var}(S_n)}} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

Then $$Z_n \to N(0,1) \quad \text{in distribution}$$

$$\left[\mathbb{P}[Z_n \geq t] \to \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx\right]$$

Chapter 1

Concentration Inequalities

- Chernoff Bound
- Sub-Gaussian RV (Hoeffding)
- Sub-Exponential RV (Bernstein)
- Johnson-Lindenstrauss
- McDiarmid inequality
- Robust Mean Estimation

Concentration Inequalities

Goal: Bounds of the form

$$\mathbb{P}\left[\frac{S_n}{n} > \mu + t\right] \leq \text{small}(n, t)$$

Prop: Let $Z \sim N(\mu, \sigma)$. Then

$$\mathbb{P}[Z \geq \mu + t] \leq e^{-t^2/2\sigma^2} \quad \forall t > 0$$

So would hope $\mathbb{P}\left[\frac{S_n}{n} > \mu + t\right] \leq C\, e^{-\frac{nt^2}{2\sigma^2}}$

Bad news from CLT:

$$\sup_t \left\{ \left| \mathbb{P}[Z_n > t] - \mathbb{P}[Z > t] \right| \right\} \text{ can be } \Omega\left(\frac{1}{\sqrt{n}}\right) \text{ even for Bernoulli}$$

where $Z \sim N(0,1)$

[See (V: 2.1) for example]

Chernoff Method: {Sub-Gaussian / Sub-Exponential} RV's

Lemma (Markov) For any non-negative X and $t > 0$, we have

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}X}{t}$$

Pf: $\mathbb{E}X = \mathbb{E}\, X \mathbb{1}_{\{X \geq t\}} + \mathbb{E}\, X \mathbb{1}_{\{X < t\}}$

$$\geq t\, \mathbb{E}\mathbb{1}_{\{X \geq t\}} + 0 = t\, \mathbb{P}[X \geq t] \qquad \square$$

Cor: (Chebycher) For any random variable X with $\mathrm{var}(X) \leq \sigma^2 < \infty$, we have

$$\mathbb{P}[\,|X - \mathbb{E}[X]| \geq t\,] \leq \frac{\sigma^2}{t^2} \quad \forall t > 0$$

More generally, suppose $\mu = \mathbb{E}X < \infty$. Then for all $\lambda \geq 0$, we have

$$\mathbb{P}[X - \mu \geq t] = \mathbb{P}[e^{\lambda(X-\mu)} \geq e^{\lambda t}] \overset{\text{Markov}}{\leq} \frac{\mathbb{E}\, e^{\lambda(X-\mu)}}{e^{\lambda t}}$$

$$\Rightarrow \log \mathbb{P}[X - \mu \geq t] \leq \inf_{\lambda \geq 0} \{ \log \mathbb{E} e^{\lambda(X-\mu)} - \lambda t \}$$

$$= -\sup_{\lambda \geq 0} \{ \lambda t - \log \mathbb{E} e^{\lambda(X-\mu)} \}$$

Define for any function $\psi: \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, the Fenchel conjugate

$$\psi^*(t) = \sup_{\lambda} \{ t\lambda - \psi(\lambda) \}$$

Let's look at the main example

$$\psi_X(\lambda) = \log \mathbb{E} e^{\lambda(X-\mu)}$$

For all $\lambda \in \mathbb{R}$, observe

$$\psi_X(\lambda) = \log \mathbb{E} e^{\lambda(X-\mu)} \geq \mathbb{E} \log e^{\lambda(X-\mu)} = 0$$

So when $\lambda < 0$ and $t \geq 0$, we have

$$\lambda t - \psi(\lambda) \leq 0 = 0 \cdot t - \psi(0).$$

Therefore for $t \geq 0$, equality holds:

$$\psi_X^*(t) = \sup_{\lambda \geq 0} \{ t\lambda - \psi(\lambda) \}$$

We therefore arrive at the bound:

Chernoff Bound:

$$\mathbb{P}[X-\mu \geq t] \leq \exp\left(-\psi_X^*(t)\right)$$

where $\psi_X(\lambda) := \log\left(\mathbb{E}\, e^{\lambda(X-\mu)}\right)$.

Ex: Let $X \sim N(\mu, \sigma^2)$. Then

$$\mathbb{E}\, e^{\lambda(X-\mu)} = e^{\frac{\sigma^2\lambda^2}{2}} \quad \forall \lambda \in \mathbb{R}.$$

So $\mathbb{P}[X \geq \mu + t] \leq e^{-\frac{t^2}{2\sigma^2}} \quad \forall t > 0$

Defn: X with mean $\mu = \mathbb{E}X$ is sub-Gaussian with parameter $\sigma > 0$ if

$$\mathbb{E}\, e^{\lambda(X-\mu)} \leq e^{\frac{\sigma^2\lambda^2}{2}} \quad \forall \lambda \in \mathbb{R}$$

If X is sub-Gaussian, so is $-X$.

$$\Rightarrow \mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

Ex: (Rademacher)
Let ε be a rademacher RV:

$$\mathbb{P}[\varepsilon = 1] = \mathbb{P}[\varepsilon = -1] = \frac{1}{2}.$$

$$\mathbb{E}[e^{\lambda\varepsilon}] = \frac{1}{2}\left(e^{-\lambda} + e^{\lambda}\right) = \cosh(\lambda)$$

Exercise: $\cosh(x) \le \exp\left(\frac{x^2}{2}\right) \quad \forall x \in \mathbb{R}$

So ε is sub-Gaussian with $\sigma = 1$.

Ex: (Bounded RV)
Suppose X is supported on $[a,b]$.
Then Jensen's inequality implies

$$\mathbb{E}\, e^{\lambda(X - \mathbb{E}_{X'}X')} \le \mathbb{E}_{X,X'}\, e^{\lambda(X-X')}$$

where X' is independent copy of X.
Observe $X - X' \sim \varepsilon(X - X')$
So

$$\mathbb{E}_{X,X'}\, e^{\lambda(X-X')} = \mathbb{E}_{X,X'}\, \mathbb{E}_{\varepsilon}\, e^{\varepsilon\cdot\lambda(X-X')}$$

$$\le \mathbb{E}_{X,X'} \exp(\lambda^2(X-X')^2/2)$$

$$\le \exp(\lambda^2(b-a)^2/2)$$

A slightly more involved argument shows that X is sub-Gaussian with parameter $\frac{b-a}{2}$

$$\Rightarrow \mathbb{P}\left[|X-\mu| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{(b-a)^2}\right)$$

Lemma: [Sum Rule]

X_i are independent
σ_i-sub-Gaussian $\forall i=1,\dots,n$ $\Rightarrow \sum_{i=1}^{n} X_i$ is $\|\sigma\|_2$-sub Gaussian

Cor: (Hoeffding)

Suppose $X_1, \dots, X_n$ are independent with $\mathbb{E}X_i = \mu_i$ and are σ_i-sub-Gaussian

Then

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i-\mu_i) \geq t\right] \leq \exp\left\{-\frac{t^2}{2\|\sigma\|_2^2}\right\}$$

$\Rightarrow$ If $\mu_i = \mu$, $\sigma_i = \sigma$ then

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu + t\right] \leq \exp\left\{\frac{-nt^2}{2\sigma^2}\right\}$$

Subexponential RV:

Ex: Let $Z \sim N(0,1)$. Let's compute

$$\mathbb{E}\left[e^{\lambda(Z^2-1)}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda(x^2-1)} e^{-x^2/2}\, dx$$

$$= \frac{e^{-\lambda}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1-2\lambda)x^2/2}\, dx$$

$$= \begin{cases} \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} & \text{if } \lambda \leq \frac{1}{2} \\ +\infty, & \text{if } \lambda > \frac{1}{2} \end{cases}$$

Defn: X with mean $\mu = \mathbb{E}X$ is subexponential with parameters (σ, α) if

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\sigma^2\lambda^2}{2}} \quad \forall\, |\lambda| \leq \frac{1}{\alpha}$$

Back to example $Z \sim N(0,1)$

$$\mathbb{E}\left[e^{\lambda(Z^2-1)}\right] \leq \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{\frac{4\lambda^2}{2}} \quad \forall\, |\lambda| < \frac{1}{4}$$

So Z^2 is $(2,4)$-subexponential.

Thm (Subexponential tail bound)
Let X be subexponential with (σ, α).
Then

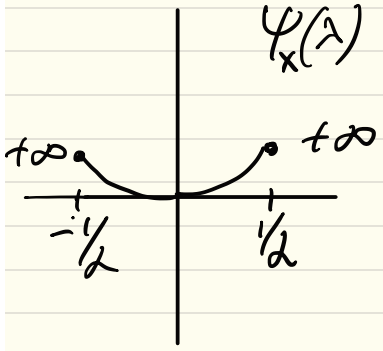
$$\mathbb{P}[X - \mu \geq t] \leq \begin{cases} e^{-\frac{t^2}{2\sigma^2}}, & \text{if } |t| \leq \frac{\sigma^2}{\alpha} \\ e^{-\frac{t}{2\alpha}}, & \text{o.w.} \end{cases}$$

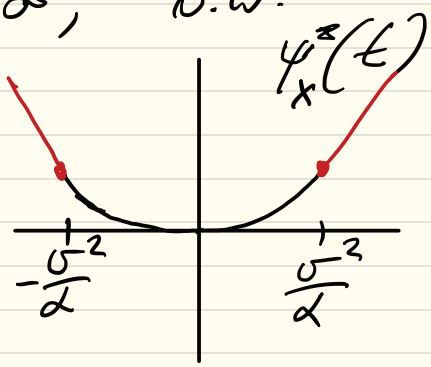
pf: Back to Chernoff

$$\log \mathbb{P}[X - \mu \geq t] \leq -\psi_X^*(t)$$

where $\psi_X(\lambda) = \log \mathbb{E} e^{\lambda(X-\mu)}$

$$= \begin{cases} \frac{\sigma^2 \lambda^2}{2}, & \text{if } |\lambda| \leq \frac{1}{\alpha} \\ +\infty, & \text{o.w.} \end{cases}$$

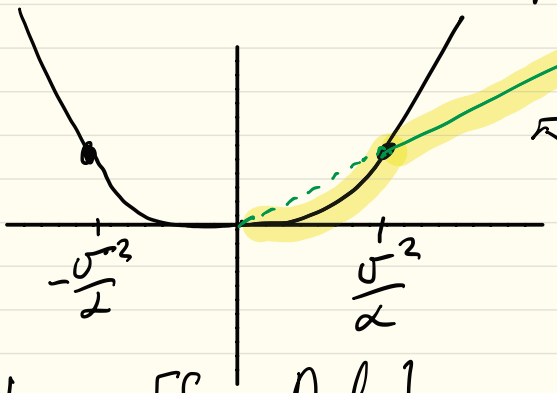




Thm: (Bernstein) Let X be subexponential with parameter (σ, α) and mean $\mu = \mathbb{E}X$. Then

$$\mathbb{P}[|X-\mu| \geq t] \leq 2\exp\left[-\left(\frac{t^2}{\sigma^2} \wedge \frac{t}{\alpha}\right)/2\right]$$

minimum



$$t \mapsto \left(\frac{t^2}{\sigma^2} \wedge \frac{t}{\alpha}\right)/2$$

Lemma: [Sum Rule]

X_i are (σ_i, α_i)-subexp $\forall i = 1, \dots, n$ $\Rightarrow \sum_{i=1}^{n} X_i$ is $(\|\sigma\|_2, \|\alpha\|_\infty)$-subexp

Thm (Bernstein for sums) Let $X_1, \dots, X_n$ be independent subexponential with parameters (σ_i, α_i), and with mean $\mu_i = \mathbb{E}X_i$. Then

$$\mathbb{P}\left[\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| > t\right] \leq 2\exp\left[-\frac{1}{2}\left(\frac{t^2}{\|\sigma\|_2^2} \wedge \frac{t}{\|\alpha\|_\infty}\right)\right]$$

Thm: (Improved Bernstein for bounded RV)
Suppose $|X-\mu| \le b$, $\mathbb{E}(X-\mu)^2 = \sigma^2$. Then

$$\mathbb{E}\, e^{\lambda(X-\mu)} \le \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right) \quad \forall\, |\lambda| < \frac{1}{b}$$

Therefore

$$\textcircled{*} \quad \mathbb{P}[|X-\mu| \ge t] \le 2e^{-\frac{t^2}{2(\sigma^2+bt)}}$$

pf: Taylor expansion:

$$\mathbb{E}\, e^{\lambda(X-\mu)} = \sum_{k=0}^{\infty} \lambda^k \frac{\mathbb{E}(X-\mu)^k}{k!} = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}(X-\mu)^k}{k!}$$

$$\le 1 + \sum_{k=2}^{\infty} \frac{\lambda^2\sigma^2}{2\cdot 3\cdot 4\cdots k} b^{k-2}\lambda^{k-2}$$

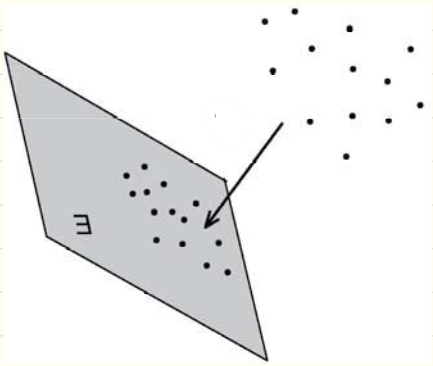
$$\le 1 + \frac{\lambda^2\sigma^2}{2}\cdot\frac{1}{1-b|\lambda|} \le \exp\left(\frac{\lambda^2\sigma^2}{2}\cdot\frac{1}{1-b|\lambda|}\right)$$

$\textcircled{*}$ Follows from Chernoff by setting $\lambda = \frac{t}{bt+\sigma^2} \in [0, \frac{1}{b})$ ∎

This is superior to Hoeffding when $\sigma^2 \ll b$.

Application: Dimensionality Reduction

Given $u_1, \ldots, u_m \in \mathbb{R}^d$ with $m \ll d$, can one map $u_1, \ldots, u_m$ into a lower dimensional space with low distortion?



Thm: (Johnson-Lindenstrauss)
Fix $\varepsilon, \delta \in (0,1)$, a set $U \subseteq \mathbb{R}^d$ of m points and a number $n > \frac{16 \ln(m^2/\delta)}{\varepsilon^2}$.
Let $X \in \mathbb{R}^{n \times d}$ consist of i.i.d $N(0,1)$ entries.
Then with probability $1-\delta$, the map $f(u) = \frac{1}{\sqrt{n}} Xu$ satisfies

$$1-\varepsilon \leq \frac{\|f(u)-f(v)\|_2^2}{\|u-v\|_2^2} \leq 1+\varepsilon \quad \forall u, v \in U.$$

pf: Observe (x_i = ith row of X)

$$\frac{\|Xu\|_2^2}{\|u\|_2^2} = \sum_{i=1}^n \underbrace{\left\langle x_i, \frac{u}{\|u\|}\right\rangle^2}_{\text{i.i.d } N(0,1)}$$

$\Rightarrow \frac{\|Xu\|_2^2}{\|u\|_2^2}$ is $(2\sqrt{n}, 4)$ – subexponential

$\Rightarrow$ Bernstein:

$$\mathbb{P}\left[\left|\frac{\|Xu\|_2^2}{n\|u\|_2^2} - 1\right| > \varepsilon\right] \le 2\exp\left[-\left(\frac{n\varepsilon^2}{8} \wedge \frac{n\varepsilon}{8}\right)\right]$$

$$= 2\exp\left(-\frac{n\varepsilon^2}{8}\right) \quad \forall\, 0 \le \varepsilon \le 1$$

So for any i, j get

$$\mathbb{P}\left[\frac{\|f(u_i - u_j)\|^2}{\|u_i - u_j\|_2^2} \notin [1-\varepsilon, 1+\varepsilon]\right] \le 2e^{-n\varepsilon^2/8}$$

Take union bound over $\binom{m}{2}$ pairs of points

$$2\binom{m}{2} e^{-n\varepsilon^2/8} \le m^2 e^{-n\varepsilon^2/8} = \delta$$ ∎

Question: What if $m = \infty$ but U has few "degrees of freedom"?

So far, we have focused on concetration of the average $\frac{1}{n}\sum_{i=1}^{n} X_i$.

Often one is interested in bounds

$$\mathbb{P}\left[\left|f(X_1,\dots,X_n) - \mathbb{E}f(X_1,\dots,X_n)\right| > t\right] \leq \text{small}(n,t)$$

where X_i are independent and $f: \mathbb{R}^n \to \mathbb{R}$ is some function.

Useful insight:

As long as $f(X_1,\dots,X_n)$ depends weakly on individual X_i, concentration holds!

Thm: (McDiarmid) Suppose that $f: \mathcal{X}^n \to \mathbb{R}$ has the bounded difference property: $\exists L_1, \dots, L_n$ such that

$$|f(x_1, \dots, x_k, \dots, x_n) - f(x_1, \dots, x_k', \dots, x_n)| \le L_k \quad \forall x, x' \in \mathbb{R}^n$$

Then for independent R.V.'s $X = (X_1, \dots, X_n)$ have

$$\mathbb{P}[|f(X) - \mathbb{E}f(X)| > t] \le 2e^{-\frac{2t^2}{\|L\|_2^2}}$$

pf: We will use the Martingale method.

Define

$$Y_0 = \mathbb{E}f(X) \quad \text{and} \quad Y_i = \mathbb{E}[f(X) | X_1, \dots, X_i] \quad \forall i$$

Observe

$$Y_i = Y_0 + \sum_{j=0}^{i-1} \underbrace{(Y_{j+1} - Y_j)}_{=: D_{j+1}} = Y_0 + \sum_{j=1}^{i} D_j$$

and

$$\mathbb{E}[Y_i | X_1, \dots, X_{i-1}] = \mathbb{E}[f(X) | X_1, \dots X_{i-1}] = Y_{i-1}$$

$$\Rightarrow \mathbb{E}[D_{j+1} | X_1, \dots, X_i] = 0$$

$$\Rightarrow \mathbb{E}[e^{\lambda(f(X) - \mathbb{E}f(X))}] = \mathbb{E}[e^{\lambda(Y_n - Y_0)}] =$$

$$\mathbb{E}\left[e^{\lambda(f(X)-\mathbb{E}f(X))}\right] = \mathbb{E}\left[e^{\lambda(Y_n - Y_0)}\right]$$

$$= \mathbb{E}\left[e^{\lambda \sum_{j=1}^{n} D_j}\right]$$

$$= \mathbb{E}\left[e^{\lambda(Y_{n-1}-Y_0)} e^{\lambda D_n}\right]$$

$$= \mathbb{E}\left[e^{\lambda(Y_{n-1}-Y_0)} \underline{\mathbb{E}\left[e^{\lambda D_n} \mid X_1, \dots, X_{n-1}\right]}\right]$$

Let X' differ from X in x_i. Then

$$\mathbb{E}\left[e^{\lambda D_i} \mid X_1, \dots, X_{i-1}\right] = \mathbb{E}\left[e^{\lambda(Y_i - Y_{i-1})} \mid X_1, \dots, X_{i-1}\right]$$

$$= \mathbb{E}\left[e^{\lambda \mathbb{E}[f(X) - f(X') \mid X_1, \dots, X_i]} \,\middle|\, X_1, \dots, X_{i-1}\right]$$

Jensen $\leq$

$$\mathbb{E}\left[e^{\lambda (f(X) - f(X'))} \,\middle|\, X_1, \dots, X_{i-1}\right]$$

$(f(X) - f(X'))$ bounded by L_i

$$\Rightarrow \mathbb{E}\left[e^{\lambda(f(X) - f(X')) \mid X_1, \dots, X_{i-1}}\right] \leq e^{\frac{\lambda^2 L_i^2}{8}}$$

So $\mathbb{E}\left[e^{\lambda(f(X) - \mathbb{E}f(X))}\right] \leq e^{\frac{\lambda^2}{8}\|L\|^2}$. Apply Chernoff.

□

Recall if $X_1, \ldots, X_n$ are independent σ-subGaussian with $\mathbb{E}X_i = \mu$, the Hoeffding implies that $\hat{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ satisfies

$$\mathbb{P}\left[|\hat{x} - \mu| \leq t \right] \geq 1 - 2\exp\left(\frac{-nt^2}{2\sigma^2}\right)$$

or equivalently

$$\mathbb{P}\left[|\hat{x} - \mu| \leq \sqrt{\frac{2\sigma^2 \ln\left(\frac{2}{\rho}\right)}{n}} \right] \geq 1 - \rho.$$

Can one achieve a similar guarantee without subGaussian assumption with a different estimator $\hat{x}$?

Answer: yes, almost!

Thm: (Median of means)

Consider $X \in \mathbb{R}$ with $\mathbb{E}X = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Let $x_1, \dots, x_n$ be i.i.d realizations of X. Subdivide into $k = 18 \log(\frac{1}{\rho})$ bins and form the empirical means $\hat{x}_j$ for $j = 1, \dots, k$. Then $\hat{x} = \mathrm{median}(\hat{x}_1, \dots, \hat{x}_k)$ satisfies

$$\mathbb{P}\left[|\hat{x} - \mu| \le \sqrt{\frac{27\sigma^2 \log(\frac{1}{\rho})}{n}} \right] \ge 1 - \rho$$

pf: By Chebychev

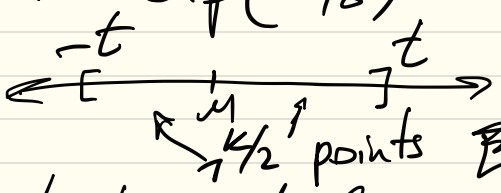
$$\mathbb{P}\left[|\hat{x}_i - \mu| \ge \sqrt{\frac{3\sigma^2 k}{2n}} \right] \le \frac{\sigma^2}{\frac{n}{k} \cdot \frac{3\sigma^2 k}{2n}} = \frac{2}{3} \quad \forall i.$$

Let $\mathbb{1}_i$ be indicator of this event. Then, by Hoeffding,

$$\mathbb{P}\left[\frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_i > \frac{1}{2} \right] \ge 1 - \exp\left(-\frac{k}{18}\right)$$

in this event,

$$\Rightarrow |\hat{x} - \mu| \le \sqrt{\frac{3\sigma^2 k}{2n}}$$



Notice that in contrast to sub-Gaussian case, $\hat{x}$ depends on confidence level ρ.