# Math 7252: High-Dimensional Statistics
# Homework 2

**due: Tuesday, November 25**

Please submit electronically directly to Canvas in a PDF file.
Each problem is worth the number of points in parentheses.
The full score is 42 points; you get "A" for 21 points, "B" for 14 points.

# 1 Covering $\ell_1$-ball with $\ell_\infty$-balls (3)

Define $B_d := \{x \in \mathbb{R}^d : \|x\|_1 \leqslant 1\}$, the unit $\ell_1$-ball in $\mathbb{R}^d$. Let $N(d, \varepsilon)$ be the minimal number of $\ell_\infty$-norm balls (also in $\mathbb{R}^d$) of radius $\varepsilon < 1$ to cover $B_d$. Show that in the regime $\varepsilon = \varepsilon(d) \in [\frac{1}{d}, \frac{2}{d}]$,

$$\log N(d, \varepsilon(d)) \asymp d \log d \text{ as } d \to \infty.$$

*Hint:* consider points with coordinates in the nodes of a regular grid. Also, you might find use for `https://en.wikipedia.org/wiki/Stars_and_bars_(combinatorics)`.

## 2 Soft thresholding (3)

Recall that in the class, we studied the Gaussian sequence model, that is, letting $\xi_t \sim \mathcal{N}(0,1)$ i.i.d.,

$$y_t = (x^*)_t + \sigma\xi_t, \quad t \in [n],$$

under the sparsity assumption $\|x^*\|_0 \leqslant s$, and analyzed the $\ell_1$-constrained estimator,

$$\widehat{x} \in \underset{\|x\|_1 \leqslant \|x^*\|_1}{\operatorname{Argmin}} \|y - x\|_2.$$

We found out that this estimator admits a "slow-rate" statistical guarantee: with probability $\geqslant 1-\delta$,

$$\|\widehat{x} - x^*\|_2^2 \lesssim \|x^*\|_1 \sigma \log(en\delta^{-1}). \tag{1}$$

Note that this guarantee is vacuous if we allow $\|x^*\|_1 \to \infty$. We now consider the penalized counterpart of this estimator, called *soft thresholding* in the literature:

$$\widehat{x} \in \underset{x\in\mathbb{R}^n}{\operatorname{Argmin}} \|y - x\|_2^2 + \lambda\|x\|_1,$$

where $\lambda$ is to be chosen. We are going to show that, for the right choice of $\lambda$, this estimator satisfies

$$\|\widehat{x} - x^*\|_2^2 \lesssim \sigma^2 s \log(en\delta^{-1}) \tag{2}$$

with probability $\geqslant 1-\delta$, which is worse than the (optimal) bound we had established for the $\ell_0$-constrained estimator in the deviation term, but the same in expectation.

1. As usual, start with the feasibility of $x^*$ in the optimization problem characterizing $\widehat{x}$:

   $$\|\widehat{x} - y\|_2^2 + \lambda\|\widehat{x}\|_1 \leqslant \|x^* - y\|_2^2 + \lambda\|x^*\|_1,$$

   whence

   $$\|\widehat{x} - y\|_2^2 \leqslant \|x^* - y\|_2^2 + \lambda(\|x^*\|_1 - \|\widehat{x}\|_1). \tag{3}$$

   Expand $\|\widehat{x} - x^*\|_2^2$ in a similar way as we did in the class, by writing $\widehat{x} - x^* = \widehat{x} - y + \sigma\xi$, expanding the square, and using feasibility to bound $\|\widehat{x} - y\|_2^2$.

2. You will have to deal with the cross term $\langle\widehat{\varepsilon}, \xi\rangle$ of the residual $\widehat{\varepsilon} = \widehat{x} - x^*$ and the noise $\xi$. To this end, we could envision using Young's inequality, writing $\langle\widehat{\varepsilon}, \xi\rangle \leqslant \|\widehat{\varepsilon}\|_1\|\xi\|_\infty$, using that $\|\xi\|_\infty = O(\sqrt{\log(en\delta^{-1})})$ w.p. $1-\delta$, estimating $\|\widehat{\varepsilon}\|_1 \leqslant \sqrt{\|\widehat{\varepsilon}\|_0}\|\varepsilon\|_2$, and concluding with the quadratic inequality to estimate $\|\widehat{\varepsilon}\|_2$ as we did in the class. Leaving aside (for the time being) the question of what to do with the other term $\|x^*\|_1 - \|\widehat{x}\|_1$, we have an issue: to get (2), we would need to have something like (putting $\delta = O(1)$)

   $$\|\widehat{\varepsilon}\|_2^2 \leqslant \sigma\sqrt{s\log n}\|\widehat{\varepsilon}\|_2 + [...],$$

   but instead of $\sqrt{s}$ we have the factor $\|\widehat{\varepsilon}\|_0$; meanwhile, $\widehat{\varepsilon} = \widehat{x} - x^*$ is *not* sparse (though $x^*$ is).

3. In fact, this issue and the bounding of $\|x^*\|_1 - \|\widehat{x}\|_1$ are handled in a unified way. Namely, use that $\|\widehat{x}\|_1 = \|\Pi_{S_*}\widehat{x}\|_1 + \|\Pi_{S_*^c}\widehat{x}\|_1$, where $\Pi_{S_*}$ is the projector on the support of $x^*$, and leverage the fact that the term $\lambda\|\widehat{x}\|_1$ in (3) is subtracted, to show that if $\lambda \geqslant \sigma\|\xi\|_\infty$, one has

   $$\|\widehat{\varepsilon}\|_2 \lesssim \lambda\sqrt{s}.$$

   Conclude with the result.

## 3 Sparse denoising with optimal rate (6)

In this problem, we improve the rate of $\ell_1$-constrained estimator to the optimal one, $s\log(n)+\log(1/\delta)$, by still using $\ell_1$-norm regularization, but changing the criterion used to fit the observations. Recall that the same rate is attained by the $\ell_0$-constrained estimator, so one might ask why is this interesting. The answer is that, as it turns out, $\ell_0$-constraint (or penalization) becomes intractable in the non-diagonal setup, i.e. sparse regression, a.k.a. $y = Ax + \sigma\xi$ with $x$ sparse; meanwhile, the new estimator generalizes nontrivially, as we will see in the lectures on estimation under shift-invariance.

- Let $u^* \in \{-1, 0, 1\}^n$ be the vector that encoded the signed support of $x^*$, i.e. $u^*_t = \mathrm{sign}(x^*_t)$.
- Note that $\|u^*\|_0 = \|x^*\|_0 \leqslant s$. On the other hand, we also know the $\ell_1$-norm of this vector,

  $$\|u^*\|_1 = s,$$

  which does not depend on $\|x^*\|_1$, and $\|u^*\|_\infty \leqslant 1$. (In fact, we know $\|u^*\|_p$ for any $p \geqslant 1$.)
- Now, let $u \odot v$ be the entrywise product of $u, v \in \mathbb{R}^n$ and consider the following estimate:

  $$\widehat{x} = \widehat{u} \odot y \quad \text{where} \quad \widehat{u} \in \operatorname*{Argmin}_{\|u\|_1 \leqslant s,\, \|u\|_\infty \leqslant 1} \|y - u \odot y\|_2^2.$$

**1°**. It turns out that this estimate satisfies the proclaimed MSE bound. Let us prove this.

(a) Decompose the MSE by expanding the square, mimicking the proof for soft thresholding:

$$\|x^* - \widehat{u} \odot y\|_2^2 = \|y - \widehat{u} \odot y\|_2^2 + [...]$$

Then, use feasibility of $u^*$ in the optimization problem defininig $\widehat{u}$.

(b) *Finish the proof. To control the stochastic terms, use Young's inequality, bound for the maximum of Gaussians, tail bound for $\chi^2$, and the following fact (no need to prove it):

*For integers $s \leqslant n$, the extreme points of the so-called polytopal ball*

$$\{u \in \mathbb{R}^n : \|u\|_1 \leqslant s, \|u\|_\infty \leqslant 1\}$$

*are $s$-sparse vectors with entries in $\{-1, 0, 1\}$.*

**2°**. Some extras:

(a) Compare with the argument for soft thresholding, and explain the improvement mechanism.

(b) Reformulate the optimization problem defining $\widehat{u}$, to obtain an explicit method for computing $\widehat{u}$ (based on sorting).

## 4 Johnson-Lindenstrauss lemma (3)

Given $n$ arbitrary points $x_1, \dots, x_n \in \mathbb{R}^d$ in a high-dimensional space, the *low-distortion embedding* (or *data sketching*) problem asks to find a mapping $F : \mathbb{R}^d \to \mathbb{R}^m$ that preserves the pairwise distances up to relative accuracy $\varepsilon \in (0, 1)$, i.e.

$$(1-\varepsilon)\|x_i - x_j\|_2 \leqslant \|F(x_i) - F(x_j)\|_2 \leqslant (1+\varepsilon)\|x_i - x_j\|_2 \quad \forall i, j \in [n]. \tag{4}$$

(This is the Euclidean distance version of the problem; one may, of course, consider other $\ell_p$-norm distances, as well as more general metric spaces.) The question is, how small one may afford $m$ to be, for a given $d, n$, and target accuracy $\varepsilon$. As it turns out, $m$ is of order $O(\varepsilon^{-2} \log n)$, regardless of $d$, and this rate is optimal for $d$ large enough. Moreover, one can choose the mapping to be linear, i.e. a matrix $A \in \mathbb{R}^{m \times d}$, and the proof is a rather straightforward application of the $\chi^2$ tail bound. *Prove the following result:*

> **JL lemma**. Let $A \in \mathbb{R}^{m \times n}$ have i.i.d. rows $a_1, \dots, a_m \sim \mathcal{N}(0, \frac{1}{m} I_d)$. There is a universal constant $c > 0$ such that for all $m \geqslant c\varepsilon^{-2} \log(2n\delta^{-1})$, (4) holds with probability $\geqslant 1 - \delta$.

In fact, you might derive the previous result from the following one:

> Let $A$ be the same as in the JL lemma. Then for an arbitrary *fixed* $x \in \mathbb{R}^d$, one has
>
> $$(1-\varepsilon)\|x\|_2 \leqslant \|Ax\|_2 \leqslant (1+\varepsilon)\|x\|_2 \tag{5}$$
>
> with probability $\geqslant 1 - \delta$, as long as $m \geqslant c\varepsilon^{-2} \log(\delta^{-1})$.

*Question:* what $m = m(\varepsilon, \delta, d)$ suffices for (5) to hold uniformly ($\forall x \in \mathbb{R}^d$ at once) with prob. $\geqslant 1 - \delta$?

## 5 RIP property for Gaussian matrices (3)

Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ have i.i.d. $\mathcal{N}(0,1)$ entries – that is, its rows $X_1, .., X_n$ are i.i.d. samples from $\mathcal{N}(0, I_d)$. Prove that for any $s \leqslant d$, $\boldsymbol{X}$ satisfies $(s, \varepsilon, \delta)$-RIP – that is, with probability $\geqslant 1 - \delta$ it holds that

$$(1-\varepsilon)\|u\|_2 \leqslant \|\boldsymbol{X}u\|_2 \leqslant (1+\varepsilon)\|u\|_2 \quad \forall u \in \mathbb{R}^n : \ \|u\|_0 \leqslant s$$

– as long as $n \gtrsim \varepsilon^{-2} s \log(ed\delta^{-1})$. In particular, a near-linear in $s$ number of samples guarantee RIP.
*Hint: you might want to solve this problem after proving the JL lemma.*

## 6 Fast JL transform (6)

This problem shares the setup with the previous one, but addresses the computational side of things.

Note that each $y_i = Ax_i$ is computed in $O(md)$, and the whole dataset is sketched in $O(nmd)$. Can we do faster than $O(md)$ per data point? In fact, there are some linear transformations of $\mathbb{R}^d$ or $\mathbb{C}^d$—i.e., corresponding to an $d \times d$ matrix $F$—that can be computed in $O(d \log d)$ for a fixed input, instead of $d^2$. If we select $m \leqslant d$ random entries of the output $Fx$, we would still be within $O(d \log d)$ time margin,[1] improving over $md$ as long as $m \gtrsim \log d$, i.e., unless sketching to a very low dimension. One such example is the Discrete Fourier transform (DFT), corresponding to

$$(F_d)_{kt} := \frac{1}{\sqrt{d}} \exp\left(-\frac{2\pi i k t}{d}\right) \quad k, t \in [d].$$

(Complex values are not the issue: one may take $Ax$ to be the real part of the submatrix of $F$.) Indeed, one may compute $Fx \in \mathbb{C}^d$ in $O(d \log d)$ via the *Fast Fourier Transform* algorithm. Similarly, one can consider the *Hadamard* matrices $\sqrt{d}H_d$, defined recursively for $d = 2^\ell$ as follows: $H_1 = (1)$,

$$H_{2^\ell} = \begin{pmatrix} H_{2^{\ell-1}} & H_{2^{\ell-1}} \\ H_{2^{\ell-1}} & -H_{2^{\ell-1}} \end{pmatrix}.$$

The advantage compared with DFT is that the entries of $H_d$ are $\pm 1$, which is condusive to applications in digital electronics/signal processing.[2]

*Explain how to compute $H_d x$, for any $x \in \mathbb{R}^d$, in $O(d \log d)$ by using recursion.*

Before we move further, I mention a seminal conjecture due to Hadamard, concerning the existence of matrices $H_d$ for $d \neq 2^\ell$. Namely, a $d \times d$ matrix is called *Hadamard of order $d$* if its entries are $\pm 1$ and its rows are mutually orthogonal. It is clear that such matrices exist for $d = 2^\ell$ (cf. the above construction, due to Sylvester); moreover, it is not hard to show that $H_d$ could only exist if $d \in \{1, 2\}$ or $d = 4k$ for $k \in \mathbb{Z}$. Hadamard's conjecture posits that $H_{4k}$ exists for all $k \in \mathbb{N}$.

**Fast JL transform.** Assuming $d = 2^\ell$, let $A = \frac{1}{\sqrt{m}} SHD$ with $D, H \in \mathbb{R}^{d \times d}$ and $S \in \mathbb{R}^{m \times d}$. Here, $D$ is a diagonal matrix with i.i.d. diagonal entries, each distributed uniformly on $\{\pm 1\}$; $H = \frac{1}{\sqrt{d}} H_d$ is the normalized Hadamard matrix; finally, $S$ samples $m$ entries of the vector at random, i.e. $S^\dagger = \begin{pmatrix} e_{k_1} & \cdots & e_{k_m} \end{pmatrix}$ where $k_1, \ldots, k_m$ are sampled without replacement from $[d]$. Note that $Ax$ is computed in $O(d \log d)$. You are now invited to reprove the following result (first published in 2009):

*For arbitrary (fixed) $x \in \mathbb{R}^d$, (5) holds w.p. $\geqslant 1-\delta$ as long as $m \gtrsim \varepsilon^{-2} \log(\delta^{-1}) \log^2(d\delta^{-1})$.*

Note that one can assume w.l.o.g. that $\|x\|_2 = 1$. The proof can be split into two parts:

**Lemma 1** (2 pts)**.** *For any $x \in \mathbb{R}^d$, with probability $\geqslant 1-\delta$ it holds that $\|HDx\|_\infty \lesssim \sqrt{\log(2d/\delta)} \|x\|_2$.*

MGF method will do the trick (or, you can refer to a suitable result from one of the first lectures).

**Lemma 2** (3 pts)**.** *Assume that $z \in \mathbb{R}^d$ is such that $\|z\|_\infty \leqslant \frac{\lambda}{\sqrt{d}} \|z\|_2$. When $m \geqslant \lambda^4 \varepsilon^{-2} \log(\delta^{-1})$,*

$$(1-\varepsilon)\|z\|_2 \leqslant \tfrac{1}{\sqrt{m}} \|Sz\|_2 \leqslant (1+\varepsilon)\|z\|_2 \quad \text{with prob. } \geqslant 1-\delta.$$

Explain how to combine the lemmas to get the result (you can do it without proving either of them).

[1] How would you sample $m$ numbers from $[d]$ uniformly at random? Suggest a method running in time $O(d+m \log d)$.

[2] In fact, one can generalize Fourier transform to (locally) compact groups. In particular, $F_d$ corresponds to the cyclic group $\mathbb{Z}_d [= \mathbb{Z} \mod d]$, and $H_{2^\ell}$ corresponds to $\mathbb{Z}_2^{\otimes \ell}$.

## 7 Gaussian width for polytopal-type sets (5)

Given a set $X \subset \mathbb{R}^d$, its *Gaussian width* is defined as

$$\mathsf{W}(X) := \mathbb{E}\left[\sup_{x \in X} \langle \xi, x \rangle\right]$$

where $\xi \sim \mathcal{N}(0, I_d)$. (Note that this definition naturally extends to sets of matrices, using the trace inner product $\langle \xi, x \rangle = \mathrm{tr}(\xi^\top x)$ where $\xi$ is a $d \times d$ matrix with independent $\mathcal{N}(0,1)$ entries.)

**1°**. Let $X$ be the unit ball of $\ell_1$-norm in $\mathbb{R}^d$. Show that

$$\mathsf{W}(X) \lesssim \sqrt{\log(ed)}.$$

**2°**. Let $X$ be the set of $s$-sparse vectors in $\mathbb{R}^d$ (assuming $s \leqslant d$) with $\ell_\infty$-norm $\leqslant 1$. Show that

$$\mathsf{W}(X) \lesssim \sqrt{s \log(ed/s)}.$$

*Hint: this calculation relies on the same fact as Problem 3: characterization of the vertices of $X$.* Compare with what we get by simply combining the previous result with the bound $\|x\|_1 \leqslant s$ on $X$.

**3°**. Let $X$ be the set of $d \times d$ matrices of rank $\leqslant r$ with operator norm $\leqslant 1$. Show that[3]

$$\mathsf{W}(X) \lesssim r\sqrt{d}.$$

**4°** Let $X$ be the set of $d \times d$ doubly stochastic matrices (i.e., with nonnegative entries, summing to 1 in each row and each column). Show that

$$\mathsf{W}(X) \lesssim d \log d + d.$$

*Hint: use Birkhoff's theorem: the vertices of the above polytope are the $d \times d$ permutation matrices.*

[3] Upon reflection, you will see that **3°** is a simple result, closer in spirit to **1°** than to **2°**. I do not know if one can improve it in a similar way to how **2°** improves over **1°**, and I doubt this...

## 8 Hypercontraction of the norm of a random vector (3)

Let $\|\xi\|_{L_p} = (\mathbb{E}[|\xi|^p])^{1/p}$. Prove that if $X \in \mathbb{R}^d$ is mean-zero and $\varkappa$-hypercontractive, i.e. one has

$$\|u^\top X\|_{L_4} \leqslant \varkappa \|u^\top X\|_{L_2} \quad \forall u \in \mathbb{S}^{d-1},$$

then the random variable $\xi = \|X\|_2$ is $\varkappa$-hypercontractive as well, i.e. one has $\|\xi\|_{L_4} \leqslant \varkappa \|\xi\|_{L_2}$.
*Hint:* start by writing $\|X\|_2^4$ as the squared sum of the squared entries of $X$.

## 9 Spectacular failure of the MGF method (4)

Let $\xi \in \{-1, 1\}^n$ be the Rademacher vector, i.e. its entries are independent (unbiased) coin tosses. In 1986, Tomaszewski conjectured that for any (fixed) vector $u$ on the unit Euclidean sphere in $\mathbb{R}^n$,

$$\mathbb{P}\left\{|\langle \xi, u\rangle| \leqslant 1\right\} \geqslant \frac{1}{2}.$$

This was proved by Keller and Klein only in 2021 (and the proof is *hardly* from the Book). In fact, proving even a much weaker statement—with *some* universal constant $p > 0$ instead of $\frac{1}{2}$—is not that easy, and things get difficult for $p > \frac{1}{3}$. See [BTNR02, Lem. A.1] for a concise proof with $p = \frac{1}{3}$.[4]

In this problem, you are *not* asked to reprove any such result. Instead, you will show that the natural strategy of proving tail bounds, via the MGF method, falls short of producing *any* result.

1. Explain why $\sup_{\|u\|_2=1} \mathbb{P}\left\{|\langle \varepsilon, u\rangle| \leqslant 1\right\} \geqslant p$ is equivalent to

$$\sup_{\|u\|_2=1} \mathbb{P}\left\{\langle \varepsilon, u\rangle \geqslant 1\right\} \leqslant \frac{1-p}{2}.$$

2. Using the MGF method, show that it would suffice to prove that

$$\sup_{\|u\|_2=1} \inf_{t \geqslant 0} \left\{ -t + \sum_{i \in [n]} \log\cosh(t|u_i|) \right\} \leqslant -\log(2) + \log(1-p).$$

3. Show the strict inequality (bummer!)

$$\sup_{\|u\|_2=1} \inf_{t \geqslant 0} \left\{ -t + \sum_{i \in [n]} \log\cosh(t|u_i|) \right\} \geqslant -\log(2).$$

[4] These issues can be circumvented if we increase the threshold from 1, but this is unsportsmanslike. More importantly, the unit threshold turns out to arise organically in certain applications, e.g., in robust optimization [BTNR02].

## 10 Concentration of sample moment tensors (6)

Here we extend the sample covariance matrix estimation result (Theorem 2.1 from Lecture 7) to higher-order moments, namely the tensor $\boldsymbol{Q}$ of 4th-order moments of $Z \in \mathbb{R}^d$. In fact, this approach is applicable to all moments; we avoid this generalization here for simplicity.

Some definitions: a quartic tensor $\boldsymbol{A} \in \mathbb{R}^{d\times d\times d\times d}$ is simply a 4-dimensional array; it is called *symmetric* if $\boldsymbol{A}_{ijkl} = \boldsymbol{A}_{\pi(i)\pi(j)\pi(k)\pi(l)}$ for any permutation $\pi$ of the multi-index. Clearly, the 4th-order moment tensor of $Z$, as given by

$$\boldsymbol{Q}_{ijkl} = \mathbb{E}\big[Z^{(i)}Z^{(j)}Z^{(k)}Z^{(l)}\big]$$

where $Z^{(i)} := \langle Z, e_i\rangle$ is the $i$th entry of $Z$, is symmetric. $\boldsymbol{A}$ is *rank-one* if $\boldsymbol{A}_{ijkl} = x_i y_j z_k w_l$ for some vectors $x, y, z, w \in \mathbb{R}^d$; in this case, one also writes $\boldsymbol{A} = x \otimes y \otimes z \otimes w$. A *symmetric* rank-one quartic tensor writes $\boldsymbol{A} = x \otimes x \otimes x \otimes x = x^{\otimes 4}$ for some $x \in \mathbb{R}^d$, and $\boldsymbol{Q}$ can be estimated from i.i.d. sample $Z_1, \dots, Z_n$ with

$$\hat{\boldsymbol{Q}}_n = \frac{1}{n}\sum_{i\in[n]} Z_i^{\otimes 4}.$$

Note that a covariance matrix is the tensor of 2nd-order moments: $\mathbb{E}[ZZ^\top] = \mathbb{E}[Z \otimes Z]$. Similarly to the case of covariance matrices, one can associate $\boldsymbol{Q}$ with a symmetric quadrilinear form that acts on a quadruple $x, y, z, w \in \mathbb{R}^d$ as follows:

$$\boldsymbol{Q}[x, y, z, w] = \sum_{i,j,k,l\in[d]} \boldsymbol{Q}_{ijkl}\, x^{(i)} y^{(j)} z^{(k)} w^{(l)}$$

where $x^{(i)} = \langle x, e_i\rangle$; in particular, $\boldsymbol{Q}[u, u, u, u]$ is a quartic form (i.e., a symmetric homogeneous polynomial of degree 4 in the entries of $u$). The *operator norm* of a symmetric quartic tensor $\boldsymbol{A}$ is

$$\|\boldsymbol{A}\| = \sup_{u\in\mathbb{S}^{d-1}} |\boldsymbol{A}[u, u, u, u]|.$$

One may show that following result for the deviations of $\hat{\boldsymbol{Q}}_n$ from $\boldsymbol{Q}$ in operator norm.

**Theorem 1.** *Assume that $Z_i \in \mathbb{R}^d$ are zero-mean and $K$-subgaussian. For $\delta \leqslant \frac{1}{n}$, with prob. $\geqslant 1-\delta$,*

$$\|\hat{\boldsymbol{Q}}_n - \boldsymbol{Q}\| \lesssim K^4\left(\frac{(d+\log(\delta^{-1}))^2}{n} + \sqrt{\frac{d+\log(\delta^{-1})}{n}}\right).$$

*In particular, the sample complexity of estimating $\boldsymbol{Q}$ up to a constant relative error in the norm is*

$$O\left(\frac{K^4}{\|\boldsymbol{Q}\|}(d+\log(\delta^{-1}))^2\right).$$

Note that $\delta \lesssim \frac{1}{n}$ is hardly a restrictive condition: it can be thought of as increasing $d$ by $\log n$.

We will prove a suboptimal version of the theorem, with $(d+\log(\delta^{-1})^3$ instead of $(d+\log(\delta^{-1})^2$. To do it, it is suggested—but not required—to follow the plan below.

1. *Approximation.* Emulating the approximation argument in the covariance estimation result, show that for any symmetric quartic tensor $\boldsymbol{A}$,

$$\|\boldsymbol{A}\| \leqslant \frac{1}{1-4\varepsilon} \sup_{u \in \mathcal{N}_\varepsilon(\mathbb{S}^{d-1})} |\boldsymbol{A}[u,u,u,u]|$$

where $\mathcal{N}_\varepsilon(\mathbb{S}^{d-1})$ is an $\varepsilon$-net of the sphere. It is OK if you get a larger universal constant than 4.

2. *Bernstein's inequality.* Take note of the following result (no need to prove it): if $W_1, \dots, W_n$ are independent random variables with $|W_i| \leqslant R$ a.s., then with probability $\geqslant 1-\delta$ one has

$$|\textstyle\sum_i W_i - \mathbb{E}[W_i]| \lesssim R \log(2\delta^{-1}) + \sqrt{\log(2\delta^{-1}) \sum_i \operatorname{Var}(W_i)}.$$

This result is proved via the MGF method; the proof mimics that of the "vanilla" $\chi^2$-bound.

3. *Truncation.* Show that if $\xi_i$ are independent with $\mathbb{E}[\xi_i] = 0$, $\operatorname{Var}[\xi_i] = 1$ and $\|\xi_i\|_{\psi_2} \leqslant K$, then

$$\left|\textstyle\sum_{i\in[n]} \xi_i^4 - \mathbb{E}[\xi_i^4]\right| \lesssim K^4 \log^3(2n\delta^{-1}) + \sqrt{n \log(2\delta^{-1})} \tag{6}$$

with probability $\geqslant 1-\delta$. To prove this result, run the truncation method as explained below.

- Define $W_i = \xi_i^4 \mathbb{1}(|\xi_i| \leqslant R^{1/4})$ and consider the decomposition

$$\sum_i \xi_i^4 - \mathbb{E}[\xi_i^4] = \sum_i (W_i - \mathbb{E}[W_i]) + \sum_i (\xi_i^4 - W_i) + \sum_i \mathbb{E}[W_i - \xi_i^4].$$

- Using the results of Exercises 3.1–3.2 from Lecture 4 (no need to prove them), show that if one selects $R \gtrsim \log^2(2n\delta^{-1})$, the right-hand side is at most $\sum_i W_i - \mathbb{E}[W_i]$ w.p. $\geqslant 1-\delta$.
- Use Bernstein's inequality (2.) to control the sum $\sum_i W_i - \mathbb{E}[W_i]$ of truncated variables.
- Control the negative deviations analogously but with some tweaks; you may assume $\delta \leqslant \frac{1}{n}$.

4. *Union bound and suboptimal result.* Combine the results of (3.) and (1.) to show a slackened version of Theorem 1 with $(d + \log(\delta^{-1}))^3$ instead of $(d + \log(\delta^{-1}))^2$.

**Remark.** Theorem 1 would follow if in (6) we manage to replace $\log^3(2n\delta^{-1})$ with $\log^2(2n\delta^{-1})$. In general, for the sum of $p$-powers under the assumptions of (3.), with any $p \geqslant 2$, one may prove that with probability $\geqslant 1-\delta$,

$$\left|\textstyle\sum_{i\in[n]} |\xi_i|^p - \mathbb{E}[|\xi_i|^p]\right| \lesssim K^p \log^{p/2}(2\delta^{-1}) + \sqrt{n \log(2\delta^{-1})}. \tag{7}$$

In particular, for $p = 2$ we recover the vanilla $\chi^2$ bound, for $p = 3$ the first term is $K^3 \log^{3/2}$, etc.; meanwhile, the truncation method, when generalized to this setting, gives $K^p \log^{\frac{p+2}{2}}(2n\delta^{-1})$, which results in $(d + \log(\delta^{-1}))^{\frac{p+2}{2}}$ for tensors. A (long) proof can be found e.g. in [HAYWC19, Thm. 3.1].

## References

[BTNR02] A. Ben-Tal, A. Nemirovski, and C. Roos. Robust solutions of uncertain quadratic and conic-quadratic problems. *SIAM Journal on Optimization*, 13(2):535–560, 2002.

[HAYWC19] B. Hao, Y. Abbasi Yadkori, Zh. Wen, and G. Cheng. Bootstrapping upper confidence bound. *In Proceedings of Neural Information Processing Systems*, 32, 2019.