ISyE 8803: Special Topics in Modern Mathematical Data Science
Homework 2

due on Sunday, 04/20 at 11:59 pm

Please submit electronically directly to Canvas in a PDF file.
Each "raw" point is worth 20 percentage points, so you can get an A by solving 3 problems.

1 Univariate exponential families and self-concordance (2pt)

By definition, *a univariate exponential family* (in canonical parameterization) is the family of p.d.f.'s

$$\left\{p_\theta(x) = e^{T(x)\theta - \phi(\theta)} \cdot \mathbb{1}_{\mathcal{X}}(x)\right\}_{\theta \in \Theta}$$

where $\Theta \subseteq \mathbb{R}$; the function $\phi(\theta)$ is called the *log-cumulant* (or *log-partition function*); $T(x)$ is *the sufficient statistic*. (Note that $p_\theta(x)$ depends on x only through $T(x)$.) An exponential family is called *regular* if the support $\mathcal{X}$ of $p_\theta(\cdot)$ is the same for all $\theta \in \mathbb{R}$. The set $\Theta^* := \mathrm{dom}(\phi)$ is the *canonical domain* of an exponential family, and the family is called *full* if $\Theta = \Theta^*$. Prove the following results:

1. The canonical domain is a convex set (i.e. segment, as $\Theta^* \subseteq \mathbb{R}$). That is, if $\theta_0, \theta_1 \in \Theta^*$, then
$$\theta_\lambda := (1-\lambda)\theta_0 + \lambda\theta_1 \in \Theta^* \quad \forall \lambda \in [0,1].$$

2. The log-cumulant is convex. (Note that it suffices to test convexity on a segment $[\theta_0, \theta_1] \subseteq \Theta^*$.)
The normalization condition
$$\int_{\mathcal{X}} p_\theta(x)dx = \int_{\mathcal{X}} \exp\left(T(x)\theta - \phi(\theta)\right) dx = 1 \quad \forall \theta \in \Theta$$
allows to express $\phi(\theta)$ as
$$\phi(\theta) = \log\left(\int_{\mathcal{X}} \exp\left(T(x)\theta\right) dx\right).$$
After that, convexity follows via Hölder's inequality, as in Problem 2 of Homework 1.

3. Let $\mathbb{E}_\theta[g(X)]$ be the expectation of $g = g(X)$ over $X \sim p_\theta$. Show that $\phi'(\theta) = \mathbb{E}_\theta[T(X)]$ and
$$\phi^{(p)}(\theta) = \mathbb{E}_\theta[(T(X) - \mathbb{E}_\theta[T(X)])^p] \quad \text{for } p \in \{2,3\}.$$
(*Hint:* to simplify calculations, you may focus on the random variable $T = T(X)$ right away.)
For the first derivative:
$$\phi'(\theta) = \frac{\int_{\mathcal{X}} T(x)e^{T(x)\theta}dx}{\int_{\mathcal{X}} e^{T(x)\theta}dx} = \int_{\mathcal{X}} T(x)\left(\frac{e^{T(x)\theta}}{\int_{\mathcal{X}} e^{T(y)\theta}dy}\right)dx = \int_{\mathcal{X}} T(x)p_\theta(x)dx = \mathbb{E}_\theta[T(X)].$$
Whence for the second derivative, using the product rule,
$$\begin{aligned}
\phi''(\theta) &= \int_{\mathcal{X}} T(x)\frac{\partial}{\partial\theta}\left(\frac{e^{T(x)\theta}}{\int_{\mathcal{X}} e^{T(y)\theta}dy}\right)dx \\
&= \int_{\mathcal{X}} T^2(x)\left(\frac{e^{T(x)\theta}}{\int_{\mathcal{X}} e^{T(y)\theta}dy}\right)dx + \int_{\mathcal{X}} T(x)e^{T(x)\theta}\frac{\partial}{\partial\theta}\left(\frac{1}{\int_{\mathcal{X}} e^{T(y)\theta}dy}\right)dx \\
&= \int_{\mathcal{X}} T^2(x)\left(\frac{e^{T(x)\theta}}{\int_{\mathcal{X}} e^{T(y)\theta}dy}\right)dx - \int_{\mathcal{X}} T(x)\frac{e^{T(x)\theta}}{\left(\int_{\mathcal{X}} e^{T(y)\theta}dy\right)^2}\left(\int_{\mathcal{X}} T(z)e^{T(z)\theta}dz\right)dx \\
&= \mathbb{E}_\theta[T(X)^2] - \mathbb{E}_\theta[T(X)]^2.
\end{aligned}$$
Note that this gives another proof of convexity. Calculation for the third derivative is omitted.

4. Construct an example showing that, in general, $\phi^{(4)}(\theta) \neq \mathbb{E}_\theta[(T(X) - \mathbb{E}_\theta[T(X)])^4]$.
(*Hint:* think in terms of familiar distributions, and Wikipedia is at your service.)

5. Let $\phi(\theta) = -\log(\theta)$ and $\mathcal{X} = \mathbb{R}_+$. Derive Θ^* and recognize the family (*hint:* take $T(X) = -X$).
Consider the family of exponential distributions $X \sim \mathrm{Exp}(\theta)$, with $\Theta = \mathbb{R}_+$, $\mathcal{X} = \mathbb{R}_+$, and

$$p_\theta(x) = \theta e^{-\theta x} \mathbb{1}_{\mathbb{R}_+}(x) = e^{-\theta x + \log\theta} \mathbb{1}_{\mathbb{R}_+}(x) = e^{-\theta x - (-\log\theta)} \mathbb{1}_{\mathbb{R}_+}(x).$$

So, this is an EF with $T(X) = -X$ and $\phi(\theta) = -\log\theta$. For the previous question, note that

$$\phi'(\theta) = -\frac{1}{\theta} = \mathbb{E}_\theta[-X], \quad \phi''(\theta) = \frac{1}{\theta^2} = \mathrm{Var}_\theta[-X], \quad \phi'''(\theta) = -\frac{2}{\theta^3} = \mathbb{E}_\theta\left[\left(\frac{1}{\theta} - X\right)^3\right];$$

here the first two central moments are well-known, and the third one can be computed by the same method as the one employed below. Now, $\phi''''(\theta) = \frac{6}{\theta^4}$, yet the fourth central moment is

$$\mathbb{E}_\theta\left[\left(\frac{1}{\theta} - X\right)^4\right] = \frac{1}{\theta^4}\mathbb{E}\left[(1-Z)^4\right] = \frac{1}{\theta^4}\sum_{k=0}^{4}(-1)^k\binom{4}{k}\mathbb{E}[Z^k] = \frac{1}{\theta^4}\sum_{k=0}^{4}(-1)^k\binom{4}{k}k! = \frac{9}{\theta^4}.$$

Here we introduced $Z \sim \mathrm{Exp}(1)$ for convenience.

- Note that for any $\theta_0 > 0$, the segment $\left\{\theta \in \mathbb{R} : (\theta - \theta_0)^2\theta_0^{-2} < 1\right\}$ is a subset of $\mathbb{R}_+$. Is that a coincidence? What is the geometric meaning of this segment in terms of function ϕ?
Not a conincidence: this set is nothing else but the Dikin ellipsoid $\Theta_{\theta_0}(1)$ of radius 1 for the cumulant $\phi(\theta)$, which is a standard self-concordant function, i.e. satisfies

$$|\phi'''(\theta)| \leqslant 2\phi''(\theta)^{3/2}.$$

As mentioned in the class, $\Theta_{\theta_0}(1) \subset \mathrm{dom}(\phi)$ for any $\theta_0 \in \mathrm{dom}(\phi)$; see [Nes13, Thm. 4.1.5].

6. Now let $\phi(\theta) = \log(1 + e^\theta)$ and $\mathcal{X} = \{0, 1\}$ (the distribution is discrete, so p.d.f. is now p.m.f.)

- Derive Θ^* and recognize the family as a reparameterized Bernoulli family.
- *Without computing ϕ'' and ϕ''' directly,* show that $|\phi'''(\theta)| \leqslant \phi''(\theta)$.
(*Hint:* use the result of 3 and compute the third moment of $X \sim \mathrm{Bernoulli}(p)$.)
Note that $\mathrm{dom}(\phi) = \mathbb{R}$. Let $X \sim \mathrm{Ber}(q)$, then

$$\begin{aligned}
p_\theta(x) = q^x(1-q)^{1-x}\mathbb{1}_{\mathcal{X}}(x) &= \exp\left(x\log(q) + (1-x)\log(1-q)\right)\mathbb{1}_{\mathcal{X}}(x) \\
&= \exp\left(x\underbrace{\log\left(\frac{q}{1-q}\right)}_{\theta} + \log(1-q)\right)\mathbb{1}_{\mathcal{X}}(x) \\
&= \exp\left(x\theta + \log\left(1 - \underbrace{\frac{e^\theta}{1+e^\theta}}_{q}\right)\right)\mathbb{1}_{\mathcal{X}}(x) \\
&= \exp\left(x\theta - \log(1+e^\theta)\right)\mathbb{1}_{\mathcal{X}}(x).
\end{aligned}$$

Here $T(X) = X \sim \mathrm{Ber}(q_\theta)$ with $q_\theta = \frac{e^\theta}{1+e^\theta} \in [0,1]$. Computing the moments we get

$$\begin{aligned}
\phi''(\theta) = \mathrm{Var}_\theta(X) &= q_\theta(1-q_\theta), \\
\phi'''(\theta) = \mathbb{E}_\theta[(X-q_\theta)^3] = q_\theta(1-q_\theta)^3 - (1-q_\theta)q_\theta^3 &= q_\theta(1-q_\theta)\left[(1-q_\theta)^2 - q_\theta^2\right] \\
&= \phi''(\theta)\left[(1-q_\theta)^2 - q_\theta^2\right],
\end{aligned}$$

and it only remains to observe that $(1-q)^2 - q^2 = 1 - 2q \in [-1,1]$ for $q \in [0,1]$.

2 Fenchel duality and generalized self-concordance (2pt)

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. Recall that the *Fenchel dual* or *convex conjugate* of f is $f_* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$,

$$f_*(u) := \sup_{x \in \mathbb{R}^d} \langle u, x \rangle - f(x). \tag{1}$$

In what follows, we assume that f is strictly convex and C^1 (continuously differentiable), and use the *involution property*: $(f_*)_* = f$. Also, you may assume $d = 1$.[1]

1°: *Maximization property.*

a. Show that f_* is differentiable at any u for which the supremum in (1) is attained, and one has

$$f'_*(u) = \arg \sup_{x \in \mathbb{R}^d} \langle u, x \rangle - f(x).$$

Use the subgradient rule for pointwise maxima of (differentiable) convex functions: "the subdifferential of the maximum is the convex combination of the gradients of active components." Let $\varphi(u, x) := \langle u, x \rangle - f(x)$. Note that $\nabla_u \varphi(u, x) = x$, whence by the active-set rule:

$$\partial f_*(u) = \mathrm{Conv}\{x(u) \in \mathrm{Argmax}_{x \in \mathbb{R}^d} \langle u, x \rangle - f(x)\}.$$

Since $f(\cdot)$ is strictly convex, the maximizer is unique, and $\partial f_*(u)$ is a singleton:

$$\nabla f_*(u) = \mathrm{argmax}_{x \in \mathbb{R}^d} \langle u, x \rangle - f(x).$$

b. Using the involution property, observe that this works in either direction, and the mappings f' and f'_* are mutually inverse (and thus bijective); in other words,

$$f'(f'_*(u)) \equiv u, \quad f'_*(f'(x)) \equiv x.$$

As such, it is convenient to define $u(x) = f'(x)$ and $x(u) = f'_*(u)$, and consider pairings $((x, u))$ with $u = u(x)$ and $x = x(u)$.

By the involution property, f is the conjugate of f_*, that is

$$f(x) = \max_{u \in \mathbb{R}^d} \underbrace{\langle x, u \rangle - f_*(u)}_{\psi(x,u)},$$

Since $\nabla_x \psi(x, u) = u$, by the active-set rule we get

$$\partial f(x) = \mathrm{Conv}\{u(x) \in \mathrm{Argmax}_{u \in \mathbb{R}^d} \langle x, u \rangle - f_*(u)\}.$$

Since $f \in C^1$, the maximizer is unique, and

$$\nabla f(x) = \mathrm{argmax}_{u \in \mathbb{R}^d} \langle x, u \rangle - f_*(u).$$

Thus, the mappings $x(u)$ and $u(x)$ are defined. To verify the mutual inverse property, note that $u(x) = \nabla f(x)$ must satisfy $x - \nabla f_*(u(x)) = 0$ by the first-order optimality condition; whence $x - \nabla f_*(\nabla f(x)) = 0$. In the same fashion one might verify that $u - \nabla f(\nabla f_*(u)) = 0$.

[1] The results can be generalized to $\mathbb{R}^d$ by fixing a segment $[x_0, x_1]$ and restricting f to $[x_0, x_1]$, i.e. defining the function $\phi(t) = f(x_t)$ on $[0, 1]$, where $x_t = (1 - t)x_0 + tx_1$. See Nesterov [Nes13] for a demonstration of this technique.

2°: *Curvature property.* Show that, in the notation defined in **1°**.b, one has

$$g''(u(x)) \equiv \frac{1}{f''(x)}, \quad f''(x(u)) \equiv \frac{1}{g''(u)}.$$

Let $g = f_*$. In dimension 1, we can use that $g' = (f')^{-1}$ and apply the inverse function rule for derivatives. To handle the general case, one might differentiate the identity $\nabla g(\nabla f(x)) \equiv x$ gives

$$\nabla^2 g(\nabla f(x)) \nabla^2 f(x) \equiv I.$$

(Actually, this uses the symmetry of $\nabla^2 f(x)$, for which we need $f \in C^2$.)

3°: *Generalized self-concordance.*

a. Assume that f is C^3-smooth and convex. Recall the definition of generalized self-concordance (GSC) with exponent $r \geqslant [1, 2]$: f is r-GSC if there exists a nonnegative constant c such that

$$|f'''(x)| \leqslant c f''(x)^r \quad \forall x \in \mathbb{R}^d.$$

For example, the "vanilla" SC function $-\log(x)$ is $\frac{3}{2}$-GSC, with $c = 2$. Prove the following:

For $r \in [1, 2]$, f is r-GSC if and only if f_* is s-GSC with $s = 3 - r$ and the same c.

Hint: use the result of **2°**.

Writing the curvature property in the form $g''(u) f''(g'(u)) \equiv 1$ and differentiating in x, we get

$$g'''(u) f''(g'(u)) + g''(u)^2 f'''(g'(u)) \equiv 0,$$

that is

$$g'''(u) = -g''(u)^2 \frac{f'''(x(u))}{f''(x(u))}.$$

Whence

$$|g'''(u)| = g''(u)^2 \frac{|f'''(x(u))|}{f''(x(u))} \leqslant c g''(u)^2 f''(x(u))^{r-1} = c g''(u)^{3-r},$$

where the last identity is by the result of **2°**.

b. Compute the dual for $g(x) = -\log(x)$ on $\mathbb{R}_+$ and $h(x) = x \ln(x) + (1 - x) \ln(1 - x)$ on $(0, 1)$. The last result is very important, we will revisit (and generalize) it in class, as *Gibbs' duality:*

Informally: entropy and log-partition function are mutual convex conjugates.

For $f(x) = -\log(x)$ one has $f'(x(u)) = -\frac{1}{x(u)} = u$, that is $x(u) = -1/u$ and

$$f_*(u) = ux(u) - f(x(u)) = -1 + \log(-1/u) = -1 - \log(-u), \quad u < 0.$$

On the other hand, for $f(x) = h(x)$, one has $h'(x) = \log(\frac{x}{1-x})$, whence

$$f'(x(u)) = \log\left(\frac{x(u)}{1 - x(u)}\right) = u$$

and $x(u) = \frac{e^u}{1+e^u}$. As the result,

$$f_*(u) = ux(u) - f(x(u)) = ux(u) - x(u) \log\left(\frac{x(u)}{1 - x(u)}\right) + \log\left(\frac{1}{1 - x(u)}\right) = \log\left(\frac{1}{1 - x(u)}\right)$$
$$= \log(1 + e^u).$$

3 Hypercontraction of the norm of a random vector (1pt)

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.

See the proof of Lemma 2.3 in the appendix of [MW17].

4 Improved union bound for the maximum of Gaussians (2pt)

Solve Exercise 3.1 from Lecture 6. You will find the definitions and context therein.
Let Q_δ be the $(1-\delta)$-percentile of $\|X\|_\infty = \max_{j\in[m]} |X_j|$, where X has the marginals $X_j \sim \mathcal{N}(0, \sigma_j^2)$ with $\sigma_j = \|a_j\|_2$. By the union bound and the standard Gaussian tail bound,

$$Q_\delta \leqslant \max_{j\in[m]} \sigma_j \sqrt{2\log\left(\frac{2}{\delta p_j}\right)},$$

for any selection of weighting probabilities $p = p_{1:m} \in \Delta_m$. By the monotonicity of $q \mapsto \frac{1}{2}q^2$ on $\mathbb{R}_+$,

$$\tfrac{1}{2}Q_\delta^2 \leqslant \max_{j\in[m]} \sigma_j^2 \log\left(\frac{2}{\delta p_j}\right).$$

Since this works for any p, one has

$$\tfrac{1}{2}Q_\delta^2 \leqslant \tfrac{1}{2}\overline{Q_\delta^2} := \min_{p\in\Delta_m} \max_{j\in[m]} \sigma_j^2 \log\left(\frac{2}{\delta p_j}\right).$$

From now on, we focus on $\frac{1}{2}\overline{Q_\delta^2}$.

1. Let $S_m := \sum_{j\in[m]} \sigma_j^2$, and observe that

   $$\hat{p}_j = \frac{\sigma_j^2}{S_m}$$

   is feasible: $\hat{p} \in \Delta_m$. Hence,

   $$\tfrac{1}{2}\overline{Q_\delta^2} \leqslant \max_{j\in[m]} \sigma_j^2 \log\left(\frac{2}{\delta \hat{p}_j}\right) = \max_{j\in[m]} \sigma_j^2 \log\left(\frac{2S_m}{\delta\sigma_j^2}\right) \tag{2}$$

   as required.

2. "Softmax inequality" states that, for any $x \in \mathbb{R}^m$ and $\beta \in \mathbb{R}_+$,

   $$\frac{1}{\beta}\log\sum_{j\in m} \exp(\beta x_j) \leqslant \log m + \max_{j\in[m]} x_j.$$

   Applying it with $\beta = 1$ and $x_j = \log(\sigma_j^2)$ we get

   $$\log(S_m) \leqslant \log m + \max_{j\in[m]} \log(\sigma_j^2) = \log(mM_m)$$

   where $M_m := \max_{j\in[m]} \sigma_j^2$. Whence

   $$\tfrac{1}{2}\overline{Q_\delta^2} \leqslant \max_{j\in[m]} \sigma_j^2 \log\left(\frac{2mM_m}{\delta\sigma_j^2}\right) \overset{(!)}{=} M_m \log\left(\frac{2mM_m}{\delta M_m}\right) = M_m \log\left(\frac{2m}{\delta}\right).$$

   where (!) is verified by simple calculus, by showing that $u \mapsto u\log(2m\delta^{-1}M/u)$ is increasing in u on $[0, M]$ as long as $2m\delta^{-1} \geqslant 1$.

3. Returning to (2), to verify that

$$\tfrac{1}{2}Q_\delta^2 = M_m \log\left(\frac{2S_m}{\delta M_m}\right)$$

as long as $\delta \leqslant 2e^{-1}$, we have to show that the sequence

$$\sigma_j^2 \log\left(\frac{2S_m}{\delta\sigma_j^2}\right), \quad j \in [m]$$

is nondecreasing under this condition. To that end, it suffices to show that $u \mapsto u\log(2\delta^{-1}S/u)$ increases on $[0, M]$ as long as $S \geqslant M$. And indeed: the derivative of this function is

$$\log(2\delta^{-1}S) - \log(u) - 1 = \log(2\delta^{-1}S/e) - \log(u) \geqslant \log(S) - \log(u) \geqslant \log(S/M) \geqslant 0. \quad \square$$

5 Orlicz norms, I (1pt)

Solve Exercises 2.1–2.2 from Lecture 4. You will find the definitions and context therein.
See the scribbled note.

6 Orlicz norms, II (2pt)

Solve Exercises 3.1–3.2 from Lecture 4. You will find the definitions and context therein.
See the scribbled note.

7 Concentration of sample moment tensors (3pt)

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 π 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 ith 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 isotropic and K-subgaussian. Then with probability $\geqslant 1 - \delta$,*

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

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

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

Remark. The second part follows from the main claim since $\|\boldsymbol{Q}\| \leqslant K^4$ (convince youself in this).

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. (Its steps can be implemented in any order, just like in our in-class proof of the sample covariance matrix result.)

1. *Approximation.* Emulating our in-class proof, show that for any symmetric quartic tensor $\boldsymbol{A}$,

$$\|\boldsymbol{A}\| \leqslant \frac{1}{1-4\varepsilon} \sup_{u \in \mathbb{S}^{d-1}} |\boldsymbol{A}[u,u,u,u]|.$$

   It is alright if instead of 4 you get another constant (but it should be a universal constant).

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

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

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

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

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

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

   - Define $W_i = \xi_i^4 \mathbb{1}(|\xi_i| \leqslant R)$ and decompose

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

   - Using the results of Exercises 3.1–3.2 from Lecture 4 (no need to prove them), show that if one selects $R \gtrsim \log^2(2n\delta^{-1})$, then the remainder sum [...] is *negative* with prob. $\geqslant 1-\delta$.
   - 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.

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 (3) 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 $\geqslant 1-\delta$,

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

In particular, for $p = 2$ we recover the vanilla χ^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

[HAYWC19] B. Hao, Y. Abbasi Yadkori, Zh. Wen, and G. Cheng. Bootstrapping upper confidence bound. *Advances in neural information processing systems*, 32, 2019.

[MW17] S. Minsker and X. Wei. Estimation of the covariance structure of heavy-tailed distributions. *arXiv preprint arXiv:1708.00502*, 2017.

[Nes13] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.