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

due on Sunday, 04/27 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 $[\theta_0, \theta_1] \subseteq \Theta^*$.)

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.)

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 $T(X) = -X$).

   - 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?

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 $X \sim \mathrm{Bernoulli}(p)$.)

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^o: *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."

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 $x = x(u)$.

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

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

3^o: *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^o**.

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.

[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 $x_t = (1 - t)x_0 + t x_1$. See Nesterov [Nes13] for a demonstration of this technique.

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.

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.
NB: updated on 04/13.

5 Orlicz norms I (1pt)

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

6 Orlicz norms II (2pt)

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

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 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 our in-class proof, 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 ε-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

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

   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

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

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

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

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.

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